
THE UNIVERSITY
*of* EDINBURGH

# Lost Pigs and Broken Genes: The search for causes of embryonic loss in the pig and the assembly of a more contiguous reference genome

Amanda Warr

A thesis submitted for the degree of Doctor of Philosophy at the University of Edinburgh
2019

Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh

# Declaration

I declare that the work contained in this thesis has been carried out, and the thesis composed, by the candidate Amanda Warr. Contributions by other individuals have been indicated throughout the thesis. These include contributions from other authors and influences of peer reviewers on manuscripts as part of the first chapter, the second chapter, and the fifth chapter. Parts of the wet lab methods including DNA extractions and Illumina sequencing were carried out by, or assistance was provided by, other researchers and this too has been indicated throughout the thesis. This work has not been submitted for any other degree or professional qualifications. All sources of information have been acknowledged.

Amanda Warr
2019

## Abstract

The pig is an economically important species, with pork being the most widely consumed meat in the world. Genomic technologies have the potential to improve reproduction, health and efficiency in the pig industry. Additionally, pigs are more similar to humans than species commonly used as medical models and improved genomic resources for the pig may facilitate its use in medical modelling. The cost of DNA sequencing has greatly decreased in recent years, allowing more researchers to incorporate next generation sequencing into their projects. Many bioinformatic tools are designed to accept a reference genome as a truth against which individuals are compared, however most of the available reference genome sequences are low-quality drafts. It is important to understand the limitations of available reference genomes in order to make full use of the sequencing technologies available. Chapter 2 assesses the quality of the published pig draft reference genome sequence, Sscrofa10.2, using short-read sequencing data from the individual from which the genome was assembled. By identifying regions where the reads disagree with the assembly, regions of low-confidence are identified and a filter is produced to reduce the impact of these regions on genomic analyses. Chapter 3 makes use of exome sequencing data to identify variants that are predicted to truncate proteins in 96 pigs, through application of filters, including the filter designed in Chapter 2. This is reduced to a short list of variants that are likely to have an impact on phenotype, specifically variants that may be associated with reproductive phenotypes and embryonic lethality. Additionally, imputation from the 96 pig

exomes to a larger set of 446 pigs each genotyped for ~60,000 single nucleotide polymorphisms (SNPs) with the PorcineSNP60 BeadChip (Illumina) is carried out, and variants are investigated for association between two reproductive phenotypes and the imputed exome variants using a genome-wide association study (GWAS) and a number of candidate genes are identified. Chapter 4 uses an alternate method of identifying phenotype altering variants by using whole genome sequencing of a trio of individuals (sire, dam and affected individual) to search for a genetic cause of foetal mummification in pigs. Finally, chapter 5 focuses on improving the available resources for the pig by reassembling the pig genome using the latest long-read sequencing technologies, producing a much improved assembly, Sscrofa11.1. The assembly is one of the most contiguous reference genomes currently available with a contig N50 of 48.2Mb, only 103 gaps remaining (less the Y chromosome) and two closed chromosomes. This project improves the available genomic resources for the pig, and identifies several putative causal variants and candidate genes underlying important traits in a commercial population.

## Lay Summary

Pork is the most widely consumed meat in the world, and it is important that we are able to produce enough pork to meet demand. In order to improve pork production, it would be beneficial to know exactly why one pig performs better or worse than others. Every living thing is made up of cells, and in those cells are a set of instructions, or “DNA”, that determine the traits of individuals along with environmental influences. We call the full set of DNA in the cell a “genome”, and the sections that specifically instruct on how to make proteins are called the “genes”. Using equipment called DNA sequencers, we can read the instructions. The DNA is written in sequences of “bases” (A, G, T and C). Most DNA sequencers can only read a few hundred bases at a time. A pig’s genome has roughly 2.6 billion bases. If we piece together sequences from the genome we can make a “reference genome”, a sequence that represents the order in which the bases are found in the complete set of DNA in an individual. Using this we can take the short sequences that DNA sequencers produce and compare them to the reference to start to understand what is different between individuals, and how those differences, or “variants”, affect the traits of the individual. For the pig, there is already a reference genome, however, in the second chapter I describe methods I used to identify parts of this reference that are badly assembled. The genome has many mistakes in it and thousands of gaps. If we compare sequences to a reference that doesn’t accurately represent any pig this can confuse analyses and make data harder to interpret. In chapter 3, I use sequences from the genes of 96 pigs and by comparing them to the

reference I try to identify variants that cause the protein to be cut short, truncating the protein potentially makes it unable to function. I use the badly assembled regions from the second chapter to exclude some inaccurate variants and identify several variants that might truncate proteins that are important for pork production. I also try to relate the presence of variants to differences in the breeding success of the pigs. In chapter 4, I look at sequences from a piglet that died during pregnancy, and both of its parents, to see if the piglet inherited any variants from its parents that may have caused its death. Each individual has two copies of a gene, one copy inherited from each parent, so even if both parents are healthy, they might each have one copy of a damaging variant, and both pass down this copy to the piglet. In chapter 2 and 3, badly assembled regions of the reference genome made accurate analysis difficult. In chapter 5, I assemble a new reference genome using a new kind of DNA sequencer that can read tens of thousands of bases at a time instead of only a few hundred. These longer sequences make it easier to find overlapping sequences and assemble the genome with fewer mistakes and fewer gaps. This new reference genome is much more accurate than the previous one and is now available for other researchers to use, and should make genome analysis in the pig more accurate and useful. This work provides other scientists with better resources to work with, and also identifies genetic variants that might cause smaller litter sizes in pigs that following further validation can be avoided to improve pork production.

## Acknowledgements

I would like to thank my primary supervisor Mick Watson for his advice, support and encouragement throughout the project. I had no experience of bioinformatics and limited programming experience when I started this project, and Mick's guidance has helped me to develop these skills. I would also like to thank the Watson group, particularly Christelle Robert who played a big part in introducing me to exome sequencing at the beginning of the project and whose data was used in chapter 3. I'd like to thank my second supervisor, Alan Archibald, who gave me the opportunity to reassemble the pig genome, this was a challenging part of the project, but also very rewarding and Alan's enthusiasm for producing a high quality reference assembly is contagious. My final supervisor, David Hume, was also very supportive and always full of ideas and advice. All three of my supervisors have shown great faith in me, and have helped to build my confidence in my abilities, for which I am extremely grateful.

My project was part funded by Genus PIC, and I'd like to thank Joseph Deeb, Matthew Campbell, Steve Rounsley and Alan Mileham for their support. Particular thanks to Joseph and Matthew, who not only advised me on science, but also arranged accommodation for me and drove me around for ~3 months while I was in the States.

Several parts of this project involved wet lab work, in which I had some experience but was not confident. I'd like to thank everyone who welcomed the poor, lost bioinformatician into their lab and were always happy to answer

silly questions and lend me equipment. From The Roslin Institute this was primarily Heather Finlayson and Christine Tait-Burkard, and from Genus ABS Matthew Campbell, Dylan Barnes, and Kimberly Kelly. Thanks to these people, and to Mick for encouraging me to pursue both, I am now much more confident working in the wet and dry lab, skills I value greatly.

I was lucky enough to do my PhD at an institute with a large number of other PhD students and postdocs, many of whom I consider good friends. We gave each other support, comic relief and a ridiculous number of pool games. While there are too many to list here, I'd particularly like to thank Alex Chambers, Omar Alfituri, Tom Marchant and William Ho.

I would like to thank my family, particularly Jenny, who isn't shy about telling people how proud she is of her little sister. Finally, my fiancé, Iain Currie, has had to spend a lot of time listening to me talk about genomes even though I suspect he isn't as interested in the subject as he pretends to be. Iain has been incredibly supportive and patient, and always reminds me not to work too hard.

# List of Abbreviations

| | |
|---|---|
| 23982 | ID number of the dam of the trio in results chapter X |
| 23982_R6 | ID number of the foetus of the trio in results chapter X |
| [N]X | N-fold coverage |
| ABS | American Breeders Service (Genus) |
| Alt | Alternative allele |
| ANKRD26 | Ankyrin Repeat Domain 26 |
| ARHGAP24 | Rho GTPase Activating Protein 24 |
| ARHGEF10 | Rho Guanine Nucleotide Exchange Factor 10 |
| ATG2B | Autophagy Related 2B |
| ATM | Amplicon Tagment Mix |
| ATP | Adenosine triphosphate |
| ATP5H | ATP Synthase Peripheral Stalk Subunit D |
| BAC | Bacterial artificial chromosome |
| BCL11B | B Cell CLL/Lymphoma 11B |
| BDKRB1 | Bradykinin Receptor B1 |
| BDKRB2 | Bradykinin Receptor B2 |
| BES | BAC end sequences |
| BLAST | Basic Local Alignment Search Tool |
| BLUP | Best linear unbiased prediction |
| BMPER | BMP Binding Endothelial Regulator |
| bp | Base Pairs |
| BR | Broad range |
| BTP | Born to purebred |
| BUSCO | Benchmarking Universal Single-Copy Orthologs |
| BWA | Burrows-Wheeler Aligner |
| BWT | Burrows-Wheeler transform |
| CALM1 | Calmodulin 1 |
| CB | ID of the sire of the trio in results chapter X |
| CCDC168 | Coiled-Coil Domain Containing 168 |
| CCDC30 | Coiled-Coil Domain Containing 30 |
| CD163 | Cluster of Differentiation 163 |
| CDKN3 | Cyclin Dependent Kinase Inhibitor 3 |
| cDNA | Complementary DNA |
| CDS | Coding sequence variant |
| CEP120 | Centrosomal Protein 120 |
| CEP295 | Centrosomal Protein 295 |
| CF | Cystic fibrosis |
| CFTR | Cystic fibrosis transmembrane conductance regulator |
| CGES | Consensus Genotyper for Exome Sequencing |
| CHAMP1 | Chromosome Alignment Maintaining Phosphoprotein 1 |
| CHORI | Children’s Hospital Oakland Research Institute |
| CLMN | Calmin |
| cm | Centimeter |
| CNV | Copy Number Variant |
| CNVR | Copy Number Variable Region |
| COL11A1 | Collagen type XI alpha 1 |

| | |
|---|---|
| CORO2B | Coronin 2B |
| CPT1A | Carnitine Palmitoyltransferase 1A |
| CRISPR | Clustered Regularly Interspaced Short Palindromic Repeats |
| CYFIP2 | Cytoplasmic FMR1 Interacting Protein 2 |
| CYP2C33 | Cytochrome P450 2C33 |
| °C | Degrees Centigrade |
| dbSNP | NCBI's database of SNPs and indels |
| DEPDC4 | DEP Domain Containing 4 |
| DHRS9 | Dehydrogenase/Reductase 9 |
| DNA | Deoxyribonucleic acid |
| dNTP | Deoxyribonucleotide tri-phosphate |
| DSTN | Destrin |
| DTNB | Dystrobrevin Beta |
| DTT | Dithiothreitol |
| EBV | Estimated Breeding Value |
| EDTA | Ethylenediaminetetraacetic acid |
| ELMO2 | Engulfment And Cell Motility 2 |
| ERLIN1 | ER Lipid Raft Associated 1 |
| ESR1 | Oestrogen Receptor 1 |
| EtOH | Ethanol |
| ExAC | The Exome Aggregation Consortium |
| EXOG | Exo/Endonuclease G |
| F | Frameshift |
| FAM216B | Family With Sequence Similarity 216 Member B |
| g | Grams (or gravitational units, see “x g”) |
| GABA | Gamma-aminobutyric acid |
| GABRA2 | Gamma-Aminobutyric Acid Type A Receptor Alpha2 Subunit |
| GABRA4 | Gamma-Aminobutyric Acid Type A Receptor Alpha4 Subunit |
| GABRG1 | Gamma-Aminobutyric Acid Type A Receptor Gamma1 Subunit |
| GATK | Genome Analysis Toolkit |
| GATK-HC | Genome Analysis Toolkit Haplotype Caller |
| GATK-UG | Genome Analysis Toolkit Unified Genotyper |
| Gb | Gigabases |
| GBS | Genotyping-By-Sequencing |
| gEBV | Genomic Estimated Breeding Value |
| GFRA2 | GDNF Family Receptor Alpha 2 |
| GLRX5 | Glutaredoxin 5 |
| GR | Glucocorticoid Receptor |
| GWAS | Genome-Wide Association Study |
| HCl | Hydrogen Chloride |
| Het | Heterozygous/Heterozygote |
| *Hin*dIII | Restriction enzyme Hind III |
| HMCN2 | Hemicentin 2 |
| Homo | Homozygous/Homozygote |

| | |
|---|---|
| HQ | High-Quality |
| HT1 | Hybridization Buffer |
| HTS | High Throughput Sequencing |
| HWE | Hardy-Weinberg Equilibrium |
| i5 | Illumina index 2 adapters |
| i7 | Illumina index 1 adapters |
| IBD | Inflammatory Bowel Disease |
| ID | Identification |
| IDY | Identity (in relation to sequence similarity) |
| IG | Immunoglobulin |
| IGF2 | Insulin-like growth factor 2 |
| IGV | Integrative Genomics Viewer |
| IL1 | Interleukin 1 |
| IL1RAP | Interleukin 1 Receptor Accessory Protein |
| IL1RN | Interleukin 1 Receptor Antagonist |
| IMGT | ImMunoGeneTics Database |
| indel | Insertion/Deletion variant |
| Kb | Kilobases |
| KDM5A | Lysine Demethylase 5A |
| LAMA4 | Laminin Subunit Alpha 4 |
| LAMB3 | Laminin Subunit Beta 3 |
| LC | Low Coverage Regions as defined by analysis in results chapter 2 |
| LD | Linkage Disequilibrium |
| lncRNA | Long Non-Coding RNA |
| LoF | Loss of Function |
| LPIN3 | Lipin 3 |
| LQ | Low Quality Regions as defined by analysis in results chapter 2 |
| LQLC | Low Quality and Low Coverage Regions as defined by analysis in results chapter 2 |
| LRRFIP2 | LRR Binding FLII Interacting Protein 2 |
| M | Molar |
| MAF | Minor Allele Frequency |
| Mb | Megabases |
| MB | Myoglobin |
| MCR | Multi-Copy Regions |
| mg | Milligram |
| miRNA | Micro RNA |
| misc_RNA | Miscellaneous RNA |
| ml | Millilitre |
| mM | Millimolar |
| mnd | Many-noded dwarfism |
| MNV | Multi-Nucleotide Variant |
| mRNA | Messenger RNA |
| MRPL58 | Mitochondrial Ribosomal Protein L58 |
| MST1R | Macrophage Stimulating 1 Receptor |
| MTFR2 | Mitochondrial Fission Regulator 2 |

| | |
|---|---|
| MTHFD1 | Methylenetetrahydrofolate Dehydrogenase, Cyclohydrolase And Formyltetrahydrofolate Synthetase 1 |
| MYOM2 | Myomesin 2 |
| N | Number of samples/Normality (unit of concentration) |
| N50 | The size at which 50% or more of the assembly is in a contig/scaffold of that size or greater |
| N90 | The size at which 90% or more of the assembly is in a contig/scaffold of that size or greater |
| NaCl | Sodium Chloride |
| NCBI | National Center for Biotechnology Information |
| ncRNA | Non-coding RNA |
| NEB | New England Biosciences |
| NFW | Nuclease Free Water |
| ng | Nanogram |
| NGS | Next Generation Sequencing |
| NHSL1 | NHS Like 1 |
| NINL | Ninein Like |
| nM | Nanomolar |
| NPM | Nextera PCR Master Mix |
| NPM1 | Nucleophosmin |
| NSB | Number Stillborn |
| NT | Neutralize Tagment Buffer |
| OLC | Overlap Layout Consensus |
| ONT | Oxford Nanopore Technologies |
| OR56B4 | Olfactory Receptor Family 56 Subfamily A Member 4 |
| OR6K6 | Olfactory Receptor Family 6 Subfamily K Member 6 |
| OTC | Ornithine carbamoyltransferase |
| PacBio | Pacific Biosciences |
| PARVB | Parvin Beta |
| PERV | Porcine Endogenous Retrovirus |
| PCR | Polymerase Chain Reaction |
| PE | Protein Existence |
| pH | Potential of hydrogen |
| PIC | Pig Improvement Company (Genus) |
| PK | Proteinase K |
| pM | Picomolar |
| PPV | Porcine Parvovirus |
| PR2 | Incorporation Buffer |
| PRKAG2 | Protein Kinase AMP-Activated Non-Catalytic Subunit Gamma 2 |
| PRRSV | Porcine Reproductive and Respiratory Syndrome virus |
| PSD4 | Pleckstrin And Sec7 Domain Containing 4 |
| PTGIS | Prostaglandin I2 Synthase |
| QC | Quality Control |
| QTL | Quantitative Trait Loci or Locus |
| RAB11FIP5 | RAB11 Family Interacting Protein 5 |
| RDM1 | RAD52 Motif Containing 1 |

| | |
|---|---|
| Ref | Reference allele |
| RH | Radiation Hybrid |
| RNA | Ribonucleic Acid |
| RPLP1 | Ribosomal Protein Lateral Stalk Subunit P1 |
| rpm | Revolutions per minute |
| rRNA | Ribosomal RNA |
| RSB | Resuspension Buffer |
| RT | Room Temperature |
| SA | Splice Acceptor Variant |
| SBS | Sequencing By Synthesis |
| SCD | Stearoyl-CoA Desaturase |
| SD | Splice Donor/Standard Deviation |
| SERPINA | Serpin Family A |
| SERPINA3 | Serpin Family A Member 3 |
| SERPINB | Serpin Family B |
| SERPINB8 | Serpin Family B Member 8 |
| SFRP5 | Secreted frizzled-related protein 5 |
| SFXN5 | Sideroflexin 5 |
| SG | Stop Gain |
| SIFT | Sorting Intolerant From Tolerant |
| SL | Stop Loss |
| SLC35C2 | Solute Carrier Family 35 Member C2 |
| SLC44A5 | Solute Carrier Family 44 Member 5 |
| SLC45A3 | Solute Carrier Family 45 Member 3 |
| SMRT | Single-Molecule Real Time |
| SNP | Single Nucleotide Polymorphism |
| SNV | Single Nucleotide Variant |
| SR | Splice Region Variant |
| SRP_RNA | Signal Recognition Particle RNA |
| SSC[1-18,X,Y] | Pig Chromosome Number |
| STOM | Stomatin |
| SUB1 | SUB1 homolog, transcriptional regulator |
| SURF6 | Surfeit 6 |
| SV | Structural Variant |
| SYNE3 | Spectrin Repeat Containing Nuclear Envelope Family Member 3 |
| TBE | Tris/Borate/EDTA |
| TCL1A | T Cell Leukemia/Lymphoma 1A |
| TCL1B | T Cell Leukemia/Lymphoma 1B |
| TD | Tagmented DNA buffer |
| TGFB2 | Transforming Growth Factor Beta 2 |
| TPMT | Thiopurine methyltransferase |
| TPT1 | Tumor Protein, Translationally-Controlled 1 |
| tRNA | Transfer RNA |
| TRRAP | Transformation/Transcription Domain Associated Protein |
| Ts/Tv | Transition/Transversion |
| TSHZ1 | Teashirt Zinc Finger Homeobox 1 |
| USDA | United States Department of Agriculture |

| | |
|---|---|
| USMARC | U.S. Meat Animal Research Center |
| UTR | Untranslated Region |
| V | Volts |
| VCAN | Versican |
| VCF | Variant Call Format |
| VEP | Variant Effect Predictor |
| VGP | Vertebrate Genome Project |
| WDR36 | WD Repeat Domain 36 |
| WES | Whole Exome Sequencing |
| WGA | Whole Genome Amplification |
| WGS | Whole Genome Sequencing |
| x g | Times Gravity |
| XIAP | X-linked inhibitor of apoptosis |
| ZEB1 | Zinc finger E-box-binding homeobox 1 |
| ZFYVE27 | Zinc Finger FYVE-Type Containing 27 |
| ZMW | Zero-Mode Waveguide |
| µg | Microgram |
| µl | Microlitre |
| µM | Micromolar |

# Contents

# Figures and Tables

# CHAPTER 1: REVIEW OF THE LITERATURE

---

*"Lisa, honey, are you saying you're never going to eat any animal again? What about bacon?"*
*"No."*
*"Ham?"*
*"No."*
*"Pork chops?!"*
*"Dad, those all come from the same animal!"*
*"Yeah right, Lisa, a wonderful, magical animal."*
-Homer and Lisa Simpson, The Simpsons

Some sections of this review have been taken, adapted and/or updated from the published review paper Exome Sequencing: Current and Future Perspectives (Warr et al., 2015a). The unedited version of this publication is in the appendix (section 7.1). The review paper was largely written by the author of this thesis, with edits made by the other authors of the review paper and minor revisions inspired by anonymous reviewers. Some sections, particularly those discussing the latest sequencing technologies, contain a small number of citations from the non-peer-reviewed, pre-print server BioRχiv owing to the fast-moving nature of advances in the field, however, this has been avoided where possible.

## 1.1 Introduction

With increasing human population size, there is a need for increased and sustainable production of dietary protein. Globally, pork is consumed more than any other meat and demand is increasing (Kristensen et al., 2014, Food and Agriculture Organization of the United Nations, 2015). Scientific research will play a leading role in the future of meat production to improve traits relating to growth, reproduction, feed efficiency, meat quality and disease resistance. Identification of loss of function (LoF) variants, including protein truncating genetic variants relevant to these traits may allow for selective breeding against undesirable phenotypes or targeted editing of these alleles to significantly reduce losses in the industry. Importantly, identifying the causative allele may allow for this to be applied to populations and species other than the one it was identified in.

In addition to the pig's importance as a food source, the species is increasingly being used in medical modelling of human diseases. Porcine models have repeatedly been found to be more similar to humans than murine models genetically, immunologically, neurologically, anatomically and metabolically (Meurens et al., 2012, Kapetanovic et al., 2012, Dolezalova et al., 2014, Swindle et al., 2011, Bassols et al., 2014). The genome of the pig has been found to have almost identical gene content to humans (Meurens et al., 2012). Transgenic porcine models have been created for a variety of human diseases including Alzheimer's Disease, Diabetes Mellitus, Cystic Fibrosis and cardiovascular disease (Fan and Lai, 2013). Identification of null mutations may allow for new models based on homozygous natural variants which can be used in biomedical research.

In order to further investigate the genomes of pigs and other livestock species, methods involving cheap and accurate high throughput sequencing (HTS) may be used. These methods allow for large cohorts of individuals to be sequenced at high depth to explore differences in their genomes. Many applications of HTS rely on the availability of a high-quality reference genome. Currently, the first step of most sequencing projects is to align reads to a reference genome of the same species, or sometimes in the case of microorganisms or species which lack a reference, a closely related species or strain (Noonan, 2010, Marston et al., 2013, Schubert et al., 2012). Errors in the reference genome will introduce errors in downstream applications and hinder the identification of important variants through an increase in poor mapping and subsequent false-positives (Zhang and Backstrom, 2014,

Phillippy et al., 2008, Salzberg and Yorke, 2005). The accuracy of a reference genome depends on several factors including the sequencing technology used for assembly and its associated errors, the depth of coverage sequenced, and the tools used to assemble, scaffold, gap-fill and annotate the assembly (Jiao and Schneeberger, 2017, Paszkiewicz and Studholme, 2010, Schatz et al., 2010, Watson, 2018). Genome assemblies are often fragmented drafts with a high likelihood of misassembly and this should be considered when using them as frameworks for analyses in re-sequencing studies (Paszkiewicz and Studholme, 2010, Salzberg and Yorke, 2005, Baker, 2012). Additionally, despite older assemblies using a highly accurate sequencing technology with relatively long read lengths and additional information such as physical maps, these were produced using older, more expensive and more laborious assembly techniques with low coverage, and are likely susceptible to errors, often remaining in a fragmented, draft state (Salzberg and Yorke, 2005). Recent advances in long-read sequencing technologies have reduced the difficulty associated with accurate genome assembly, particularly with their ability to span large repeat regions, which cannot be accomplished with short reads. Current long-read sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have a much lower accuracy than Illumina's short-read sequencing technology and older Sanger sequencing technology.

This review will discuss sequencing technologies and methods for the characterisation of large complex genomes; methods of genome

interrogation focusing on the use of whole genome sequencing (WGS) and whole exome sequencing (WES) for the identification of variants; examples of the uses of WES in a range of species; and the potential to implement these techniques in pigs to find variants that impact economically important phenotypes.

## 1.2 A Brief History of Genome Assembly

The first sequencing of the human genome took more than 200 scientists over a decade in a project that cost almost $3 billion before it was declared complete (International Human Genome Sequencing Consortium, 2004), and even with additional resources that have been allocated to this effort since then and much improved technology, the human assembly is still not truly complete (Liu et al., 2014). Over the past decade the price of genome sequencing has plummeted through the use of high-throughput sequencing (Mardis, 2008) and so-called next generation sequencing (NGS) technologies, much of the work has become automated and methods have improved.

Historically, the assembly of genomes has not been a trivial task and has proven a challenge both to sequencing technology and to bioinformatics (Bao et al., 2011, Ekblom and Wolf, 2014, Horner et al., 2010, Jiao and Schneeberger, 2017, Schatz et al., 2010, Treangen and Salzberg, 2012). There is currently no sequencing technology that can demonstrably sequence whole chromosomes in a single fragment, it is therefore necessary to use bioinformatic techniques to assemble genomes from high coverage

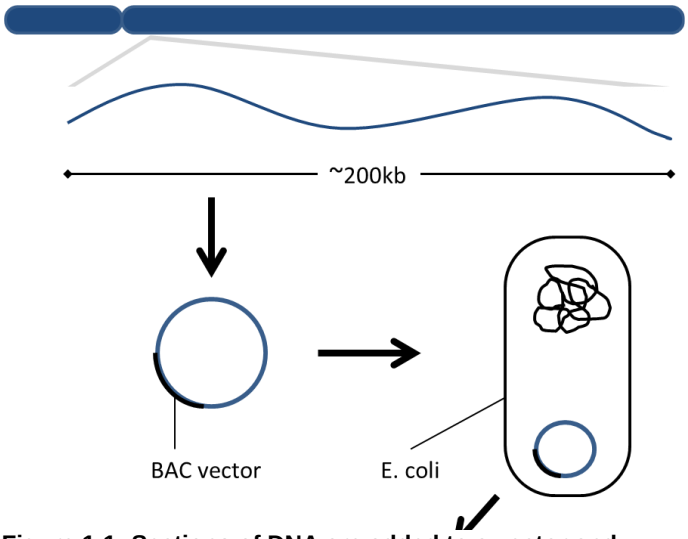



**Figure 1-1- Sections of DNA are added to a vector and grown in E.coli as Bacterial Artificial Chromosomes (BACs)**

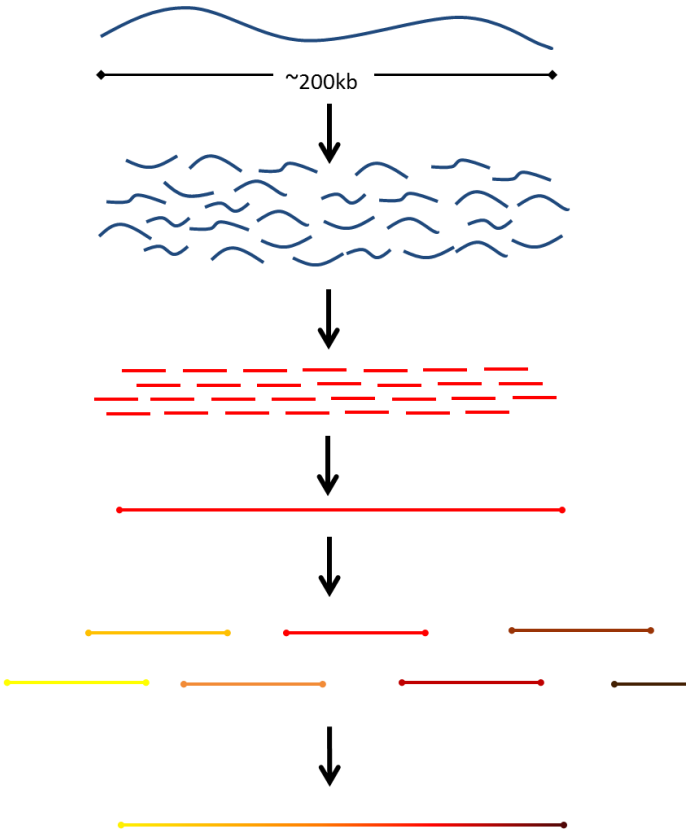



**Figure 1-2- BAC-by-BAC assembly involves the breaking down of genomes into 200kb segments, which are further fragmented before sequencing. These sequences must be assembled first into the BAC they came from and then into the genome. DNA fragments are shown in blue, sequence in yellow, red and brown.**

sequencing data by identifying regions of sequences from smaller fragments that are highly similar over a length sufficient to suggest they originate from overlapping regions of the genome. Early methods of genome assembly employed bacterial artificial chromosomes (BAC) where DNA fragments from the species of interest a couple of hundred kilobases in length are inserted into the F-plasmids of bacteria which are grown to amplify the fragment (figure 1-1) before Sanger sequencing. In this hierarchical shotgun sequencing approach the sequenced fragments are then assembled individually and subsequently assembled together (figure 1-2; International Human Genome Sequencing Consortium, 2004, Groenen et al., 2012, Osoegawa et al., 2004). While BAC-by-BAC assembly typically employs Sanger sequencing, which is a highly accurate sequencing technology, it is expensive, labour-intensive, uses low-coverage sequencing and can be difficult to assemble both within individual BACs and between BACs. Whole genome assembly was initially confined to large consortia assembling the human genome and the genomes of important medical models due to the costs and time required for this method.

Over the last decade there have been major advances in sequencing technology, reducing the cost and increasing throughput (Schatz et al., 2010, Watson, 2014). These advances have allowed more researchers to sequence a wider range of species at higher depth. The advancements have also brought new bioinformatic challenges for assembly, with most HTS fragments being much smaller than Sanger sequenced fragments of BACs, initially just tens of bases in length, but now hundreds of bases in length (Bao

et al., 2011). As a result it is much more difficult to have confidence that highly similar sequences originated from the same genomic location. This problem is made more challenging by HTS having fluctuating coverage from sequencing bias causing some regions to be underrepresented, and by the presence of repetitive elements and low-complexity regions in genomes (de Koning et al., 2011, Haubold and Wiehe, 2006) which can only be assembled confidently by sequence fragments that span their length (Baker, 2012, Treangen and Salzberg, 2012, Roberts et al., 2013, Berlin et al., 2015). While high-throughput, short read sequencing has made genome assembly a much cheaper process, this has proven to be a trade-off against the quality of the analysis and assembly of genome sequences and has produced genomes that are highly fragmented (Baker, 2012, Salzberg and Yorke, 2005, Hernandez et al., 2008). In addition to actually generating and assembling sequences, the assembled contigs (or contiguous sequence) must be scaffolded and assigned to chromosomes. In the past, scaffolding has been done using pairs of short sequences representing the ends of fragments of known size such as the end sequences from fragments cloned in BAC or fosmid libraries but also uncloned fragments from so-called mate pair libraries. Assignments to chromosomes have been done using physical mapping techniques such as fluorescent *in situ* hybridization and radiation hybrid mapping. Most researchers, however, do not have these resources available for their species of interest and sequence is not always reliably assigned to chromosomes, if at all.

More recently, sequencing methods that produce long reads have been developed and are a popular choice for genome assembly as they can be assembled into highly contiguous sequences (contigs) and span large regions that cannot be assembled by shorter reads (Roberts et al., 2013, Jain et al., 2018a, English et al., 2012). The longer contigs produced make assigning sequence to chromosomes easier in species where there is some knowledge of gene positions, and a small number of scaffolds can often be obtained per chromosome. Again, this advance introduced new bioinformatic challenges with an even higher error rate than the previous generation of sequencers, but with the advantage that relatively few regions cannot be incorporated into contiguous sequence (Berlin et al., 2015).

## 1.3 Important Developments in Sequencing Technologies and Associated Tools for Genome Assembly

### 1.3.1 Sequencing technologies

There are a number of platforms used for short read sequencing including a range of Illumina (CA, USA; formerly Solexa, Cambridge, UK) platforms, Ion Torrent (Life Technologies, CA, USA) and 454 Pyrosequencing (454 Life Sciences/Roche, Switzerland; discontinued). Short read technologies are based on sequencing by synthesis (SBS) and generally involve fixing DNA molecules in place in a nanowell, or on a surface or bead with nucleotides and polymerases in solution. The DNA is amplified (usually through PCR) to form a cluster of identical strands. By denaturing the DNA and synthesising a

complementary strand, the base incorporated can be detected in a way that differs by technology: coloured fluorescent tags in Illumina, natural release of hydrogen atoms during base incorporation in Ion Torrent, and a flash of light from luciferase in 454. These technologies have a number of restrictions and flaws, notably most involve a PCR step to amplify each strand so that base incorporation is detectable. PCR creates bias as it is less reliable in GC- or AT- rich regions (Shin et al., 2013, Aird et al., 2011, Chen et al., 2013) and these may be underrepresented in the final product. Illumina now offer methods using polymerases with less extreme GC-bias and library preps that are PCR free to overcome these problems, and these have been found to perform better in these regions (Rhodes et al., 2014, Williams et al., 2012). Perhaps the most important limitation of these platforms is the read length. The addition of bases in waves is not entirely efficient and strands from the same cluster can skip ahead or lag behind which creates noise in the signal, the more sequencing cycles pass, the more strands will become out of sync until the true signal is completely obscured, this is called phase error (Fuller et al., 2009). While 454 sequencing has fallen out of use, Illumina and Ion Torrent remain, with Illumina being the most widely used and accurate short read HTS technology available. Recently Illumina sequencing has been improved in the HiSeq X series through the introduction of ordered flow cells in which clusters are formed in nanowells. These clusters are more densely packed and lead to more usable data per flow cell than clusters formed on a flat surface.

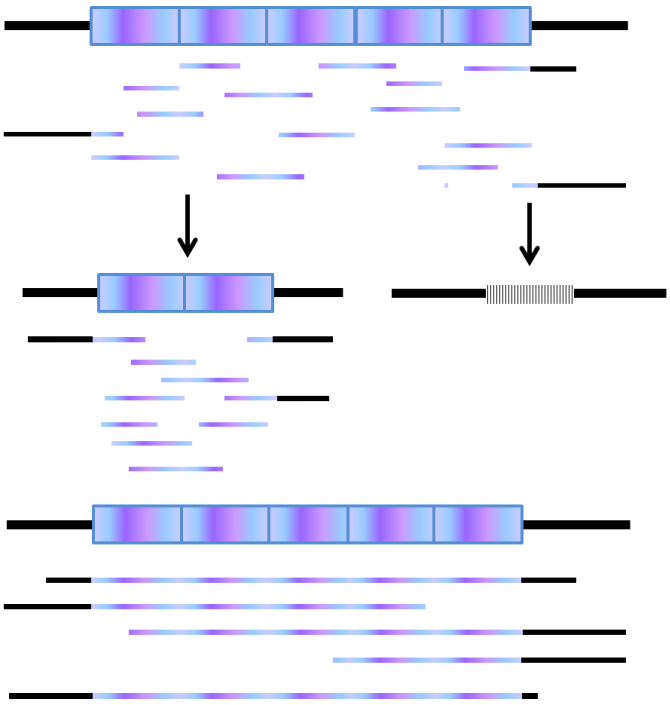


**Figure 1-3- In areas with tandem repeats (represented by coloured blocks), short reads (top) will often produce either a collapsed repeat or an unresolved gap (grey). Long reads that are able to span the entire repeat region and reach into flanking regions may assemble these correctly (bottom).**

While they have high throughput and are accurate on the sequence level, the short length of these reads make them a poor choice for genome assemblies as they cannot span low-complexity regions and repeats. Assemblies using only short-reads will often fragment in low-complexity or repeat regions (figure 1-3; Roberts et al., 2013).

Recently, new technologies have emerged that overcome many of the problems encountered by short read sequencing technologies. Long read sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) do not need to amplify DNA and do not suffer from phase error, they also sequence in real-time, not requiring cycling waves of base incorporations. While currently short read sequencing from Illumina remains the HTS technology with the highest sequence accuracy, long read sequencing can span complexities in genomes and allow for much more complete and contiguous assemblies.

In PacBio sequencing, the polymerase rather than the DNA is held in a fixed position in a well called a zero-mode waveguide (ZMW) which holds a small enough volume that cleaved fluorescent tags can be detected from base incorporation. The double stranded DNA molecule has a hairpin with a recognisable sequence ligated to each end and when a nucleotide in the hairpin interacts with the fixed polymerase sequencing is initiated. Fluorescently tagged nucleotides in solution are incorporated by the polymerase into the DNA strand and the tag is cleaved, detected and diffuses out of the ZMW. Owing to the hairpins on each end of the strand and depending on the length of the strand, each strand can be read multiple

times allowing for a consensus sequence with increased accuracy, this is limited by the functional life time of the polymerase and for longer fragments will include fewer passes (Rhoads and Au, 2015). One of the major limitations of PacBio sequencing is its error rate of 11-15% (Rhoads and Au, 2015, Korlach, 2015). While errors in PacBio data are often stated to be randomly distributed, they are more common in homopolymers. Errors are likely caused by failure of the fluorescent tags to diffuse quickly enough, tags diffusing too quickly to properly detect or dissociation of a nucleotide from the active site before phosphodiester bonds have formed, resulting in indel errors (Eid et al., 2009). This means that for PacBio reads to be accurate there must be redundancy in the sequencing (Chin et al., 2013). However, this need for redundancy is common to most sequencing strategies and technologies, including the highly accurate Illumina technology and is evident in the need for high sequence coverage of the target genome during assembly. For example, the DISCOVAR strategy for assembling draft genome sequences based solely on Illumina data requires sixty-fold genome coverage (Weisenfeld et al., 2014). If fragment lengths are long, as is desirable in genome assembly, there may not be a sufficient number of passes over the sequence to correct the read before the polymerase fails. Consequently, high coverage or the use of more accurate short reads is necessary to correct the reads. Additionally, due to the indel errors being more common in homopolymers, these regions tend to be inaccurate even after error correction. For certain applications, such as Iso-Seq which is a PacBio protocol for RNA sequencing, there are usually a sufficient number of

passes over a single fragment to attain a high accuracy sequence without the use of alternate technologies to correct them (Sharon et al., 2013).

ONT's sequencing method differs from those mentioned so far in that it does not involve a polymerase incorporating bases and is the only sequencing technology to detect the template DNA sequence itself and not base incorporation during complement synthesis (Loman and Watson, 2015). In ONT sequencing a pore is fixed in a membrane and an ionic current is run through the pore, the DNA is guided through the pore by the electrophoretic effect of the current. A motor protein restricts the speed that DNA can pass through the pore and disruptions in the current are measured as the DNA passes through the pore, the bases can be determined by the pattern of disruption to the current, or "the squiggle". ONT has produced the longest read length of any of the sequencers, with their "ultra-long" reads recently exceeding 2 Mb (Payne et al., 2018). There are several forms of library preparation for the platform that have different accuracies, read lengths, presence or absence of incorporation of the compliment sequence and protocol lengths for different applications. ONT's MinION sequencer is the smallest and cheapest sequencer available and has received much attention for its portability, being used during disease outbreaks (Hoenen et al., 2016, Quick et al., 2017, Quick et al., 2016), in the Antarctic and down mines for environmental samples (Johnson et al., 2017, Edwards et al., 2017) and on the International Space Station (Castro-Wallace et al., 2016), to name a few. Similarly to PacBio, ONT has a higher error rate than short read technologies, at 15%, and these are largely random, but occur non-randomly

in homopolymeric regions as these cause the squiggle to remain constant and currently the number of bases must be determined by the length of time the signal is constant (Lu et al., 2016). The majority of errors can be corrected with a consensus from high coverage and ONT predict that those remaining in homopolymers will be corrected through improvements in base-calling and modifications to the pore. Thus far, ONT have not been major competitors to PacBio and Illumina owing to lower throughput and higher error rates, however with accuracy improving, lower costs and the availability of new, larger sequencers GridION (up to 5 flow cells) and the more efficient PromethION (up to 48 flow cells), ONT sequencing is becoming increasingly popular in the field of genome assembly with a number of genomes being assembled using this technology (Jansen et al., 2017, Loman et al., 2015, Jain et al., 2018a, Risse et al., 2015). However, currently PacBio is still the more commonly used of the two for large genome assembly.

### 1.3.2 Genome assembly tools

Both ONT and PacBio have presented new challenges for bioinformatics tools through the length of their reads and their higher error rates. Most bioinformatics tools before these technologies were optimised for short and highly accurate Illumina data and tools have had to be modified and designed anew for sequence base calling, read correction, assembly, sequence alignment and variant calling that account for the differences in the technologies.

There are a large number of genome assembly tools for short read sequencing including Velvet (Zerbino and Birney, 2008), SOAPdenovo2 (Luo

et al., 2012), ABySS (Simpson et al., 2009), Opera (Gao et al., 2011) and Celera (Myers et al., 2000). Tools for long-read assemblies include those which assemble using only long reads, and hybrid assemblers that make use of both long and short reads. Assembly tools for long read sequencing include Canu (Koren et al., 2017), Falcon (Chin et al., 2016), Tulipa-julia (Formerly Tulip; Jansen et al., 2017) and Miniasm (Li, 2016).

Assembly tools fall into four major classes of assembly paradigms: overlap-layout-consensus (OLC), string graph, greedy assemblers and de Bruijn-graph (Nagarajan and Pop, 2013). OLC assemblers identify overlapping sequence using an all-against-all alignment, create a layout graph based on the alignment with a node for every read and an edge between reads that overlap, and create a consensus sequence for the contig (Pop, 2009). De Bruijn graphs break the reads into *k-mers* and link the *k-mer* nodes based on the order in the reads, assembling the genome by finding the most likely path through the nodes based on the links provided by the reads (Pevzner et al., 2001). De Bruijn graphs rely on accurate sequence, so long reads must be corrected before assembly with this approach (Nagarajan and Pop, 2013). String graphs are very similar to OLCs, but they remove redundant information from the graph, reducing the computational resources required (Nagarajan and Pop, 2013, Myers, 2005). Unlike de Bruijn graphs, string graphs do not reduce the assembly to *k-mers* but can have nodes of any size allowing them to retain more information. Greedy assemblers assemble based on the alignment that has the greatest immediate benefit beginning with the most well aligned sequences and extending from there, if the

placement of the read does not contradict already placed sequence, it will be placed (Pop, 2009). Greedy assemblers are prone to misassembly and while early assemblers used this paradigm, they have largely been replaced by more selective methods. OLC can be very computationally intensive, particularly with high-depth short-reads, as each read essentially acts as a node, whereas with de Bruijn graphs, in theory the number of nodes is equal to the genome size regardless of depth of coverage, however sequence errors will introduce novel false *k-mers* and these will increase with coverage. De Bruijn graphs also have benefits over OLC by not carrying out a computationally intensive alignment step, and from the fact that the *k-mers* already contain consensus information (Li et al., 2012, Nagarajan and Pop, 2013). Depending on *k-mer* size and repetitive content of the genome being assembled, de Bruijn graphs are more likely to struggle with crossing repeat regions as the sequence has been broken down into smaller chunks that will not span them. De Bruijn -graphs may also struggle to separate similar sequences from multiple loci across the genome (Nagarajan and Pop, 2013). String graphs handle repeats better as they do not break the reads down into *k-mers* and in fact the nodes are often longer than the reads having been formed from overlaps between reads, better incorporating repeats into the assembly (Myers, 2005). The successful assembly of repeat regions in a string graph depends on the ability of the sequencing technology to span them. On the computational side of things, string graphs have solved the problem of genome assembly. The limiting factor for genome assembly now

is the length of repeats in the genome being assembled and the ability of the sequencing technology to span these.

Following assembly of contigs, they ideally need to be ordered and oriented with respect to the position they fall on the chromosome, this process is called scaffolding. In the past, methods such as radiation hybrid mapping, linkage mapping, physical mapping and fluorescence *in-situ* hybridisation have been used to order and orient contigs and assign them to chromosomes. These methods are labour-intensive and expensive. Computational methods are much cheaper and more widely used at present, though often do not successfully create chromosome-assigned assemblies without additional data.

Assembly tools often have their own associated scaffolder, though standalone scaffolders are also available. A common method of scaffolding is to use long reads or paired reads such as BAC end libraries, fosmid libraries, mate pair libraries or paired end libraries. Paired reads are reads from the ends of a molecule whose length is known approximately and where the orientation of the reads relative to each other is known. Where one read aligns with one contig (contig A) and its paired read aligns with a different contig (contig B) then contig A can be linked and oriented relative to contig B (figure 1-4). The data is aligned with the contigs and string graph or greedy algorithm is used to find likely joins between contigs. For the graph-based scaffolders, the contigs are the nodes and the joining reads provide the edge information. As the insert size of a paired-end library is usually fairly constant,

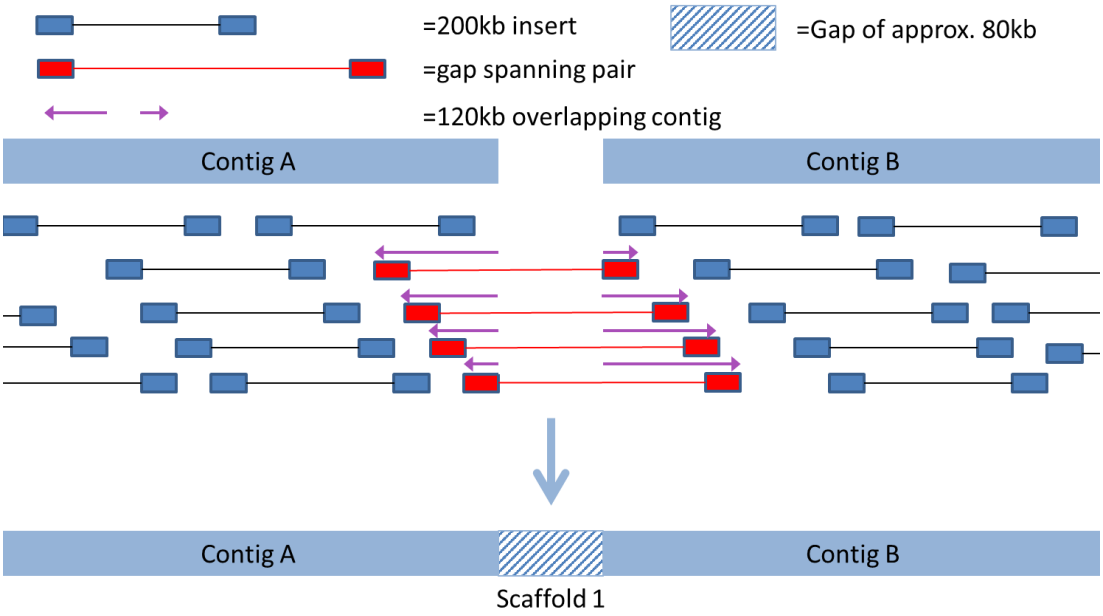



**Figure 1-4- Paired reads, when aligned to contigs, may align with one of the pair each on separate contigs (red), as the approximate distance and orientation relative to each other is known, the two contigs can be placed adjacently in the correct orientation with a gap between them of an estimated size. Here approximately 120kb of contigs are spanned by a pair that originated from a fragment of approximately 200kb, so the gap size can be estimated at 80kb.**

the gap sizes between the contigs can be estimated reasonably accurately. The use of paired read data can be effective, but in cases where the gap is larger than the library's insert size, the gap cannot be closed. An example of a greedy assembler is SSPACE (Boetzer et al., 2011) and examples of graph based assemblers are ABySS (Simpson et al., 2009), AHA (Bashir et al., 2012), BESST (Sahlin et al., 2014), GRASS (Gritsenko et al., 2012), MIP (Salmela et al., 2011), Opera (Gao et al., 2011), SCARPA (Donmez and Brudno, 2013), SGA (Simpson and Durbin, 2011), SOAPdenovo2 (Luo et al., 2012) and SOPRA (Dayarian et al., 2010). Scaffolders like these may be more effective with end sequences from libraries with very large insert sizes such as BAC libraries. The data used for assembly may not be sufficient to close the gaps, particularly in regions where the gaps exist due to low coverage. Although the price of sequencing is dropping, scientists working on species that are not directly related to human health and food security may not be able to afford extra sequencing to close these gaps. A universal set of BACs for scaffolding, designed based on knowledge of evolutionary break points, has been put together for birds (Damas et al., 2017) with plans to do the same for mammals, while this does risk introducing false joins, these can potentially be identified and corrected later in the process.

An alternative method of scaffolding is to align the contigs with the genomes of one or more closely related species and attempt to order and orient the contigs based on these genomes. This can be done with or without additional paired end data. As the number of available reference genomes increases, this becomes a more attractive option. However, the method relies on the

accuracy of the reference genome(s) and the relationship between the species used. Even the most closely related species can have evolutionary break points that will complicate this process and increase the potential for false joins and as such it is fairly error prone. Examples of tools that use this method are MeDuSa (Bosi et al., 2015), RACA (Kim et al., 2013), Ragout (Kolmogorov et al., 2014), CONTIGuator (Galardini et al., 2011) and scaffold_builder (Silva et al., 2013).

The quantity and context of gaps remaining in the assembly after scaffolding will depend on the sequencing technology, depth of sequencing and assembly method used (Hunt et al., 2014). Assemblies using only short-reads will have a large number of gaps, particularly in repetitive regions that cannot be spanned by the length of the read. Other gap-causing features of short-read data include sequence that appears in multiple loci across the genome and fluctuating coverage often caused by GC-bias. The number of gaps can be reduced through the use of gap fillers. Scaffolding adds information to the process of extending contigs by identifying the sequence that sits either side of the gap (figure 1-5), which can help to find the specific sequence that fits there using either long reads that span the gap, through assembling long reads that reach into the gap, or using paired-end and mate-pair data in short-reads to assemble the missing sequence. Examples of tools for gap filling are PBJelly (English et al., 2012), GapFiller (Nadalin et al., 2012), GapCloser-SOAPdenovo2 (Luo et al., 2012), and Sealer (Paulino et al., 2015).

With the high error rate of long-reads, additional tools are needed for error-correction before and after assembly. Some of these tools use only the long reads and correct by aligning the reads to themselves or to assembled contigs and taking a consensus sequence for example, Quiver (Chin et al., 2013), Arrow (Unpublished), Racon (Vaser et al., 2017) and Nanopolish

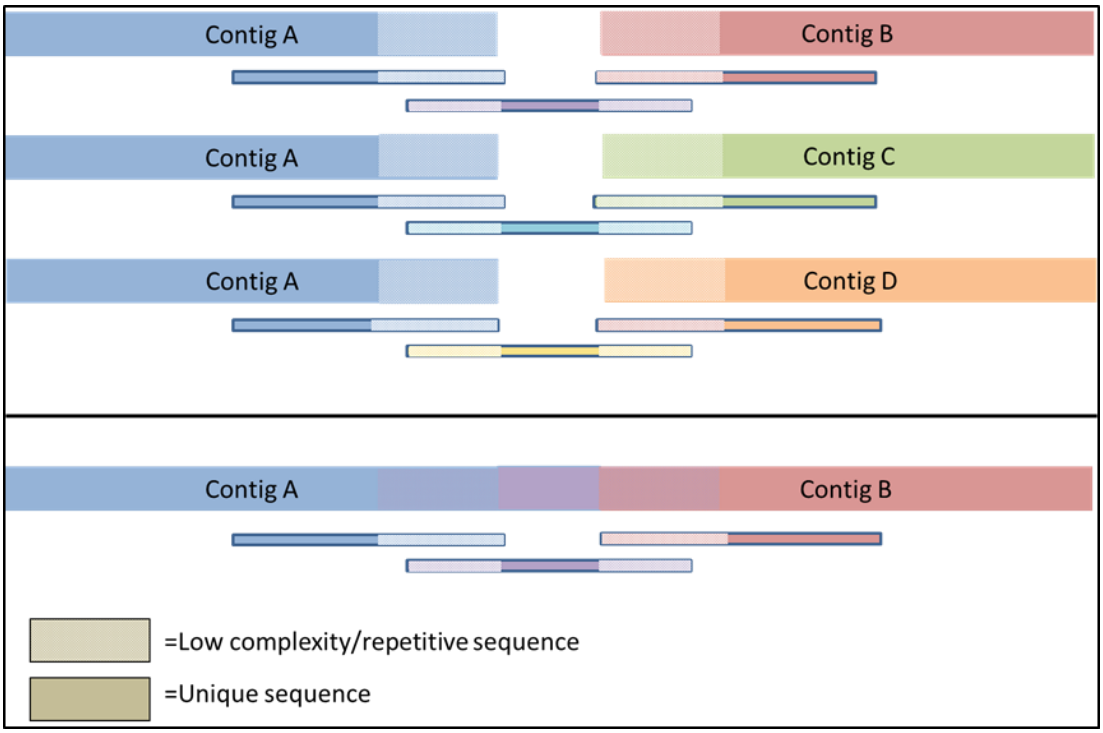



**Figure 1-5- During assembly a gap may be caused by there being multiple possibilities for which contigs belong together, using alternative methods to scaffold can provide enough information to close the gap with a gap filler. In this case, once it is known that Contig B follows Contig A, the sequence that belongs in the gap can be identified.**

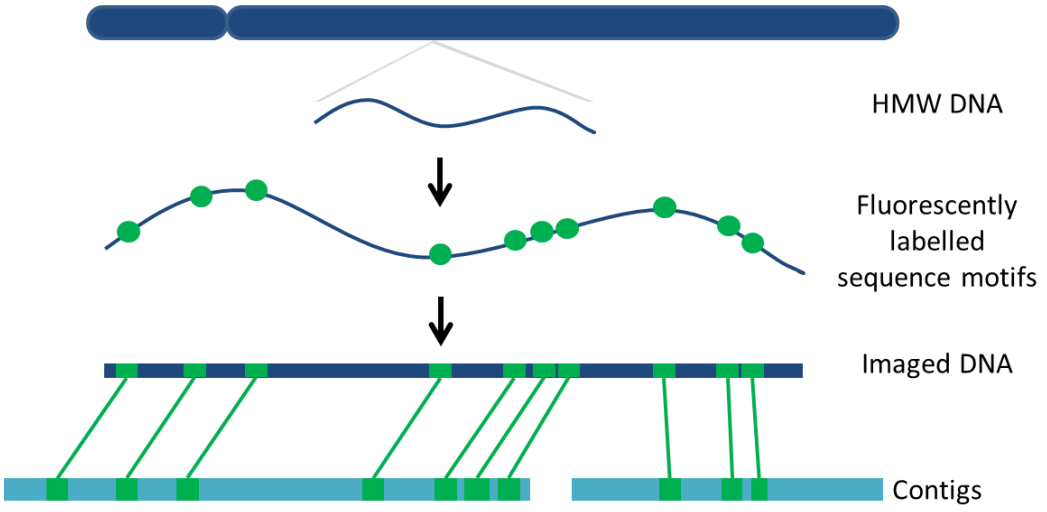



**Figure 1-6- Long DNA fragments can have specific motifs labelled and visualised. The same motif can be identified in the assembled contigs to create a kind of bar code that helps determine the correct order of the contigs and estimate gap size. This is the process used by Bionano.**

(Loman et al., 2015). Others align highly accurate short reads to the long reads or contigs and correct them based on the more accurate reads, for example Pilon (Walker et al., 2014). The use of accurate short reads to correct the assembly is self-evidently limited by the ability to map the short reads to unique locations in the assembly. Thus, it is likely that correction of repetitive sequences will be less effective. These tools essentially behave like variant callers, altering the reference where there are high-confidence homozygous SNPs or indels.

### 1.3.3 Additional methods for scaffolding and improving contiguity

The problem of scaffolding has been significantly decreased by long-read sequencing as there are far fewer contigs following initial assembly and the contigs are longer. Several companies now offer additional specialised sequencing with various methods to improve the contiguity of contigs from long reads and improve scaffolding, a few of these will be briefly discussed. Optical mapping using systems such as Bionano use enzymes to nick DNA at specific sequence motifs and hybridise fluorophore-labeled probes at the location of the nick, the DNA stand is then imaged in a nano-channel and the pattern of flourophores can be recorded along with approximate distances (figure 1-6). The “barcode” of the motif from the imaging can then be matched up with the locations the motif occurs in sequenced contigs to order and orient the contigs (Das et al., 2010). Phase Genomics offers Hi-C sequencing (Belton et al., 2012, Lieberman-Aiden et al., 2009), a method that can reveal the three dimensional structure of folded DNA in the cell.

Formaldehyde is used to create cross links between chromatin while the DNA is still in the cell. The DNA is then fragmented with restriction enzymes and sticky ends are closed with biotinylated and thiolated probes, the crosslinks are reversed, and biotinylated fragments are pulled down. The biotinylated probes are then removed, and the fragments are sequenced using an Illumina sequencer. When the fragments are sequenced they yield paired ends that originate from varying distances apart on the strand, however the more frequently two sequences appear linked, the more likely they are to have been in close proximity to one another in the cell. This allows both for scaffolding and for identification of biologically relevant physical proximity of different regions of the genome in the cell. Dovetail use a method similar to Hi-C sequencing. In a Dovetail Chicago library, chromatin is reconstituted upon naked DNA, instead of using naturally occurring chromatin. This provides similar scaffolding information, but removes any connections caused by biologically relevant physical proximity of DNA in order to reduce the associated noise from such interactions. Dovetail's HiRise pipeline can then scaffold contigs based on the linkage information from these reads and the probability that sequences were in close proximity (Putnam et al., 2016). 10X offer artificial long reads which can be used for assembly, scaffolding and phasing based on methods by Zheng et al. (2016). These long reads are created by separating the source DNA into droplets containing a relatively small amount of the genome each (figure 1-7), where each droplet has a unique sequence tag which is attached to the DNA during barcoding. Following sequencing and alignment to contigs, the unique

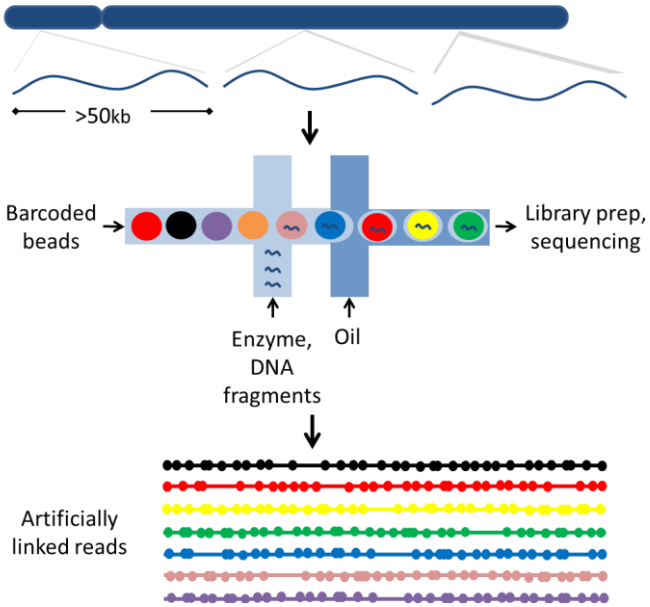



**Figure 1-7- The genome is fragmented into pieces >50kb in size and each fragment is placed in an individual droplet, fragmented and barcoded. Following Illumina sequencing the barcodes allow reads from the same location in the genome to be identified.**

sequence tag allows reads that originated from the same location to be identified and used to scaffold contigs (Kitzman, 2016). The 10X Genomics approach can also be used for assembly, essentially employing a divide and conquer approach that breaks the assembly of a complex genome into a series of smaller and more tractable assembly challenges reminiscent of the BAC-by-BAC approach but at much greater depth.

## 1.4 The Future of Genome Assembly

A number of projects have been set up to increase the number of species with high quality reference genomes, including the Vertebrate Genome Project (Vertebrate Genomes Project, 2018), the Bird 10,000 Genomes project (Zhang, 2015) and the Oz Mammal Genomes project (Oz Mammal Genomes, 2016). These projects provide recommendations of technologies and methods that should be used to sequence, assemble and annotate genomes to get the best quality reference assemblies. The Vertebrate Genome Project, for example, currently recommend PacBio sequencing at 60X or greater, assembly using Falcon unzip (Chin et al., 2016) and error correction with Arrow. For scaffolding they recommend Bionano with at least 2 enzymes at 80X coverage and/or 10X linked reads. They recommend this is followed with further scaffolding using Hi-C at 100X. Gap filling should be carried out using PBJelly (English et al., 2012). Finally, the assembly should be polished using Illumina with at least 50X coverage. The Vertebrate Genome Project aims to sequence ~66,000 genomes based on taxonomic

hierarchy, beginning with 266 genomes (one genome from each order) in phase 1.

While reference genomes such as these will likely be invaluable to biological research, as the currently available genomes have been, the concept is imperfect. It was recognised before the first human genome was sequenced that treating a single individual as a reference for a variable population is a flawed approach (Walsh and Marks, 1986). For many species, assembling a genome from a single individual is standard, even though the assembly does not truly represent that individual. Most reference genomes represent haploid genomes of diploid or polyploid species, meaning that they often switch back and forth between haplotypes and consequently do not truly represent the genome of any single individual or haplotype (figure 1-8). Additionally, reference genomes such as the human genome (International Human Genome Sequencing Consortium, 2004) and genomes of microorganisms do not represent a single individual, with reference genomes incorporating all sequence across a range of individuals in a single linear reference. Incorporating sections of sequence into a linear genome that do not consistently occur in all individuals in the species or population, such as race-specific regions in humans, cause the reference to be fragmented, however, leaving this sequence out prevents complete analysis of individuals carrying this additional sequence (Liu et al., 2014).

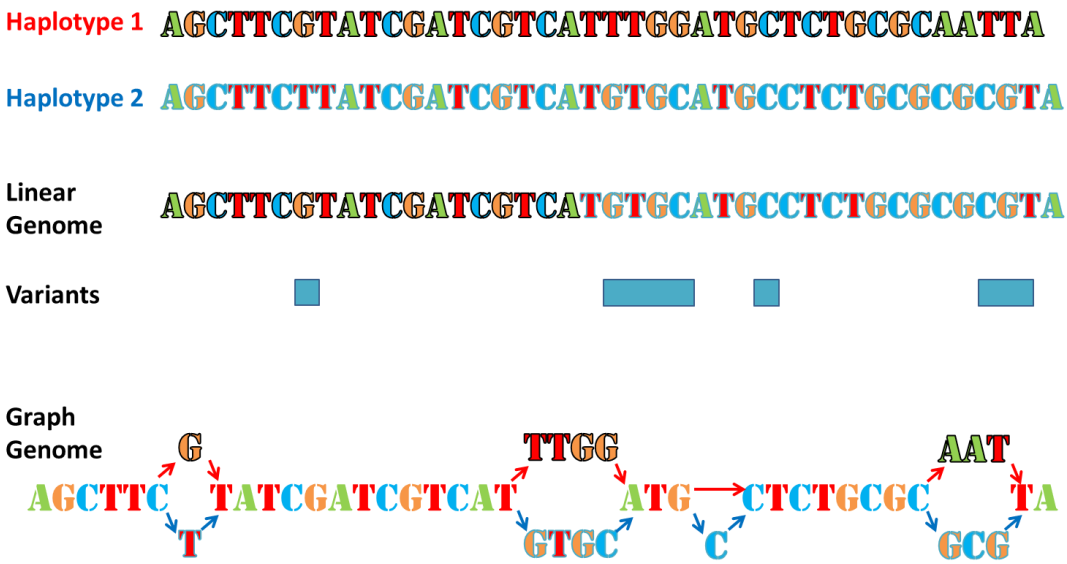



**Figure 1-8- Linear reference genomes assembled from diploid species often flip back and forth between the haplotypes without accurately representing either individual. Graph reference genomes incorporate both haplotypes.**

In mice, multiple assemblies are being used to represent different lines within the species (Keane et al., 2011, Yalcin et al., 2012a, Yalcin et al., 2012b), and similarly multiple assemblies are now being produced for cattle. Currently, however, there are still many species lacking a single high-quality reference and the teams such as Ensembl that make reference genomes accessible as annotated sequences in public domain genome browsers are reluctant to accept multiple references for a single species. To better incorporate alternative sequences into an assembly without causing fragmentation or the need for multiple linear assemblies, graph based reference genomes (figure 1-8) have been discussed (Myers, 2005, Compeau et al., 2011, Church et al., 2015, Church et al., 2011). New reference formats present new challenges for methods of genome visualisation, annotation and analysis. While these formats will better represent the species, commonly used analysis tools cannot currently accept them and the addition of multiple haplotypes across the genome may increase compute time for already time-consuming parts of analyses such as alignment and variant calling. For these to become more commonly used, the community will need to put considerable effort into creating and validating new tools and agree on file formats, this will likely delay their uptake. One of the benefits of linear genomes is the fixed co-ordinate system that allows for simple recording and visualisation of additional data such as gene annotation. The visualisation of graph genomes in particular would be difficult. Currently linear genomes are presented as one or more sequences with “N”s indicating gaps where the assembly is incomplete, visualisation

tools are designed to display genomes with a variety of annotation tracks (Gundersen et al., 2011). If a simple way of visualising graph genomes and associated information cannot be designed, this may discourage their use.

New methods have been used to create phased, diploid genome assemblies of diploid species using long reads to better represent the haplotypes present in the individual being sequenced (Chin et al., 2016). Recently advances have been made in the phasing of large genomes in individuals that are hybrids of two highly divergent breeds. Hi-C sequencing of the individual allows the two divergent haplotypes to be largely separated, effectively producing two haploid, phased assemblies, one representing the haplotype received in the gamete from each parental breed. This has recently been incorporated into the Falcon assembly pipeline as an optional module and is known as Falcon-Phase (Kronenberg et al., 2018). Trio binning can also be used without Hi-C data, where the F1 individual and both of its parents are sequenced without additional scaffolding information, the F1 animal using long reads and the parents using short reads. The haplotypes can be effectively separated using only this data by identifying k-mers from each parental genome in the short read data and associating the long reads with these. The effectiveness of this method depends on the heterozygosity of the F1 animal (Koren et al., 2018). Traditionally, individuals that were particularly inbred were preferred for genome assembly to try to reduce heterozygosity as much as possible and reduce the impact of switching from one haplotype to another in the linear assembly. In the future, it is likely more F1 hybrid

individuals will be sequenced and haplotype resolving methods employed to establish two haploid genomes.

Furthermore, with advances in assembly techniques and long read sequencing length, accuracy, throughput, and affordability, some speculate that reference genomes will become obsolete with genomes being assembled de-novo on a project-by-project basis. While this is technically possible, particularly for species with smaller, less complex genomes, this comes with high costs and a computational burden both in terms of carrying out the assembly and in storing and analysing the data, with additional challenges in terms of annotation and the lack of fixed coordinates which will impact on communication of results. For extremely large genomes (Leitch et al., 2010) and for repetitive and/or polyploid genomes, as is the case in many plant genomes such as the wheat genome (hexaploid, >50% repetitive content; Garbus et al. (2015)), this is extremely unlikely to be feasible any time soon. Another option, however, is reference-guided assembly. By aligning relatively low coverage long-read data and Illumina WGS data to the species' reference genome, small and large structural and base-level variation can be found and the reference genome can be modified to represent a new individual (Schatz, 2018; https://github.com/schatzlab/crossstitch). It is not entirely clear how well this method would handle large translocations or spans of sequence that are present in the new individual, but absent in the reference, which would essentially require a small *de-novo* assembly. However, this may be a much more cost-effective way to produce multiple genomes.

## 1.5 The Pig Genome

The importance of the pig in agriculture and as a biomedical model has driven a large amount of research on the pig genome (Groenen, 2016). Previously, linkage maps have been used to aid identification of quantitative trait loci (QTL) in the pig genome (Archibald et al., 1995, Rohrer et al., 1996), these maps provided researchers with a number of microsatellites to use in mapping experiments for important traits (Rathje et al., 1997). Subsequently, chips for genotyping thousands of single nucleotide polymorphisms (SNPs) were developed (Ramos et al., 2009) and exploited in genome-wide association studies (GWAS) to map QTL. To date 26,076 QTL representing 647 different traits and published in 601 papers are recorded for pigs in the Animal Quantitative Trait Loci Database (AnimalQTLdb; https://www.animalgenome.org/cgi-bin/QTLdb/index, accessed 19th April 2018). Research using comparative mapping suggested a high degree of similarity between the pig and human genomes, though with gene order variations between the two species (Sun et al., 1999). This degree of similarity is not observed between humans and mice.

Sanger sequencing technologies were applied to the pig genome with an initial assembly using 0.66X coverage (Wernersson et al., 2005), this assembly further supported the previously observed similarities between human and pig sequences and suggested a high similarity in isochore structure. More recently the Swine Genome Sequencing Consortium (Schook et al., 2005) published the Sscrofa10.2 draft reference genome assembly of a Duroc sow using BAC-by-BAC assembly techniques similar to those used to

complete the first human genome (Groenen et al., 2012). While the assembly used the best tools available at the time, evidence suggests there are a number of errors (Warr et al., 2015b) including an unexpectedly large number of structural variations occurring across all individuals in studies (Paudel et al., 2013) and likely misplaced sequence creating false signals in genome wide association studies (van Son et al., 2017, Yang et al., 2013). In addition to not being entirely accurate this assembly still has gaps, portions of sequence on unplaced scaffolds and portions of sequence completely missing. However, it has allowed for extensive analysis of the genome and, with the continued decrease in sequencing costs, many resequencing studies have been carried out that would not be possible without the assembly and associated annotation. In addition to the publication of the assembly, analyses by Groenen et al. (2012) provided evidence of genomic changes involved in the evolution of the suid lineage; evidence of at least two independent centres of domestication in Europe and Asia, respectively; and signatures of selection in the genome. Another genome assembly of a Chinese breed, the Wuzhishan miniature pig has also been produced (Fang et al., 2012). This assembly is said to more closely resemble other Chinese breeds than the Duroc breed assembly does, it is however even more fragmented than Sscrofa10.2 as it was assembled using Illumina short-reads.

Recognising that the reference genome is incomplete, Li et al. (2017a) sequenced nine pig breeds (Hampshire, Berkshire, Landrace, Pietrain, Large White, Bameim Jinhua, Meishan and Rongchang) using 100X Illumina sequencing and assembled them using SOAPdenovo. While these

assemblies are highly fragmented, the authors were able to identify over 1000 genes in each assembly that were missing in Sscrofa10.2 including 871 that were assembled in all nine genomes and the Tibetan wild boar genome (Li et al., 2013a) and are likely erroneously missing from Sscrofa10.2.

The use of genomic data, currently largely in the form of SNP chips, is becoming very common in pig breeding, and owing to the relatively short generation interval and the high number of individuals per litter, has the potential to increase the frequency of beneficial variants and purge detrimental variants at a rapid rate. The use of technologies such as SNP chips relies on variants from a stretch of the genome being inherited together, or linkage disequilibrium (LD). The level of LD depends on a number of factors relating to population history including the size of the founding population, the level of inbreeding and the level of migration. LD is also influenced by natural and artificial selection, genetic drift and effective population size. LD is not constant across the genome and is affected by patterns of recombination. Studies have demonstrated that there are genomic features that influence the rate of recombination, these include GC content, repeat elements, gene density, level of gene expression, nucleosome formation and epigenetic modification. Recombination rates also vary by position on the chromosome, sequence complexity, location of specific sequence motifs and the gender of the individual (Tortereau et al., 2012, Paigen and Petkov, 2010, Myers et al., 2008, Kong et al., 2002). In pigs, recombination rates have been shown to be higher in females than males, with the exception of chromosome 1 which is likely due to a large AT

rich region on the chromosome. Recombination in females is substantially reduced in AT rich regions, with males being less sensitive to this sequence context (Tortereau et al., 2012, Meunier and Duret, 2004). Regardless of gender, the highest recombination rate in pigs occurs on chromosome 12, with the lowest occurring on chromosome 1 (Tortereau et al., 2012). As has been found in other species, recombination hot spots in pigs tend to cluster around the chromosome end regardless of the position of the centromere (Tortereau et al., 2012). There is evidence in domestic breeds of increased LD in regions under strong selection for production traits, particularly where previous studies have reported QTL and genes that may be associated with important production traits (Amaral et al., 2008, Nsengimana et al., 2004).

The different population histories of Chinese and European breeds are evident in the differing level of LD between these groups. Chinese breeds tend to have lower LD than European breeds (Amaral et al., 2008, Ai et al., 2013), which is consistent with Chinese breeds being more diverse than European breeds. This likely ties in with both the differing breeding and selection practices between Western and Asian farms, and the differing sizes of the founding populations (Amaral et al., 2008). European wild boar have LD between that of domesticated European and Chinese breeds, this relatively high LD in wild boars is likely due to a population bottleneck that caused extinction of wild boars in the British Isles and Northern Europe leading to higher levels of inbreeding in the wild populations (Amaral et al., 2008). Amaral et al. (2008) reported a maximum haploblock size of 400kb in European breeds and 10kb in Chinese breeds. Similarly, Veroneze et al.

(2013) found that in six commercial pig lines the average block size was ~395kb with blocks of 100-400kb being most common in all lines. Studies have found that the rate of LD decay in pigs is slow, which is beneficial to selective breeding efforts as SNP associations with causative variants will break down less often, reducing the need for frequent marker reassessment and allowing for the use of relatively low density SNP chips (Amaral et al., 2008, Akanno et al., 2014, Badke et al., 2012, Veroneze et al., 2013).

## 1.6 Interrogating a Genome

### 1.6.1 Why use genomics in animal breeding?

Historically in agriculture, selection for improvement has centred around breeding high performing individuals together to produce high performing offspring (Hill, 2014). The problem with this method, particularly in livestock, is that it is very time consuming and expensive. For this method the animal must be kept until the phenotype of interest can be measured, with the potential of wasting money rearing poor performers. One way around this is to use an estimated breeding value (EBV), using known data of the performance of the parents and siblings to estimate the performance of the individual using best linear unbiased prediction (BLUP; Henderson, 1975). Usually an EBV incorporates a number of different phenotypes and may include data from individuals in different environments. This method is faster than traditional methods, however with no information on the underlying causes of differences in phenotype it is not entirely efficient. In specific cases where a genetic cause or strong association with a marker is known, these specific loci can be tested using PCR-based or restriction fragment length

polymorphism methods for application in marker assisted selection (Muir, 2007, Southern, 1982, Saiki et al., 1988). Genome-wide estimated breeding values (gEBV) are now popular among animal breeders. These breeding values are based on an array of genotyped loci spread across the genome and phenotype data may be used both for selection for markers associated with traits and for discovery of new associations (Meuwissen and Goddard, 1996, Meuwissen et al., 2001, Xu, 2003, ter Braak et al., 2005, Goddard et al., 2016). While the causative variant is often not known, the information gained from interrogating the genome is more accurate than EBV and sufficient to bring about a rapid change in the phenotype of a population.

### 1.6.2 SNP chips

In animal breeding the most common method of interrogating the genome is using a SNP chip (Ramos et al., 2009, Meuwissen et al., 2001, Matukumalli et al., 2009). SNP chips are arrays designed to identify the genotype of specific SNPs spread across the genome. Oligos that target the alleles of biallelic SNPs are fixed on an array, and DNA of the animal to be tested is introduced, the DNA will preferentially bind to the oligo that matches its genotype and thus by measuring the DNA binding to each oligo, the genotype can be predicted with high accuracy (Ragoussis, 2009). Sometimes the SNPs on a chip are specific SNPs known to be relevant to a trait, but the majority are SNPs spread at roughly equal distances across the whole genome (Ramos et al., 2009). Though it is unlikely that these SNPs are causative of a phenotype themselves unless they have been specifically selected for inclusion on a chip for this reason, it is possible to find a

causative region based on the LD between the tested SNP and the causative variant. Depending on how strongly a trait is being selected for and how inbred a population is, associating a phenotype with these SNPs can point to very large regions of the genome, as more SNPs will be in LD with causative variants. Additionally, complex traits are influenced by multiple genes across the genome and rarely just one specific variant, this leads to multiple quantitative trait loci (QTLs) being associated with a trait with the causative variants of each not being identified. It is not necessary to know the exact causative variant for selective breeding, however, as long as the tested SNP and the variant remain in LD, the causative variant will be selected for/against as well (Villumsen et al., 2009). One problem with this is that it is likely that this connection between the SNP and the cause only apply to the population in which the haplotype is circulating and cannot necessarily be applied to other populations, additionally, it is possible for the SNP and the causative variant to become disassociated over time through recombination. Knowing what the causative variant is allows for stronger and more reliable selection of a phenotype and may also allow for a better understanding of the mechanism behind how the variant causes the change in phenotype. Identification of the causative gene also has the potential of allowing the information to be applied to other species, which may be of particular interest in medical models.

### 1.6.3 Sequencing

With the falling price of sequencing (Beckmann, 2015, Drmanac, 2011, Snyder et al., 2010), it is now possible to affordably shotgun sequence whole

genomes with short reads. By comparing these sequences with the reference genome of the species, or a closely related species, it is possible to identify SNPs, indels and some large structural variants, some of which may be causative of a trait. One of the limiting factors of this method is that large cohorts may be needed to find specific variants and, particularly in species with large genomes, WGS is still very expensive. WGS data takes a lot of computer power to process and a large amount of disc space to be stored (Xuan et al., 2013). In species with a high level of repetitive content in the genome, much of these data and associated processing are essentially useless. It is difficult or often impossible to accurately map short-reads in repetitive regions. Additionally, the function of the majority of the genome is poorly understood and analyses tend to focus on protein coding genes. For studies involving small numbers of individuals, WGS offers rich data, however for large cohorts the cost and resources required remain prohibitive.

An alternative method is exome sequencing. The exome traditionally comprises of all exons of protein coding genes and typically covers between 1% and 2% of the full size of the genome, depending on the species. The target region may also be extended to target functional non-protein coding elements (e.g. miRNA, lncRNA etc.) as well as specific candidate loci. This can be done by using probes designed to capture the target region and enrich the sample for the fragments containing target sequence. Following exome capture, the fragments must be sequenced, which is usually done using a short-read sequencing platform in a similar manner to WGS. While focussing on the exons loses information from regulatory regions it allows for

sequencing of the most well understood regions of the genome. Variants can be identified in these regions that are likely to have a direct impact on the function of proteins and potentially phenotypes. Variants associated with phenotypes in these regions that are not expected to be causative may be in LD with causative variants that are outside of the exome and can be used in a similar manner to SNP chip variants to select against detrimental phenotypes. However, the ultimate goal is to find the direct cause of different phenotypes which should allow for the most efficient selection against detrimental phenotypes.

## 1.7 Alignment tools

Once sequencing data are obtained, mapping is the first step for most bioinformatic pipelines and often has the longest runtime. The aim of mapping is to find the original genomic location of each read relative to a reference genome. This must be done as accurately as possible while allowing for true variation and sequencing errors, but must be balanced with the reasonable use of time and resources. Importantly, this process relies on the accuracy of the reference genome. Errors in the reference genome cause inaccurate mapping and false positives in downstream analysis (Salzberg and Yorke, 2005, Phillippy et al., 2008, Warr et al., 2015b).

For most of the available mapping tools the first step is to create an index which can be searched, this can either be an index of the genome, an index of the reads or both (Li and Homer, 2010). Hash table-based tools were the first to be used on short-reads from NGS and involve making a hash table

index of *k-mers* which can then be searched for the locations of *k-mers* in the query sequence. The query sequence is usually a seed consisting of an adjustable number of bases from the 5' end of the read which is the most accurate end of the read in short-reads as the 3' end is subject to phase errors. Most hash table-based aligners now use gapped seeds in which only a certain number of positions in the seed need to match the *k-mer* allowing for a specified number of mismatches and indels in the seed to account for sequencing error and true variants in the DNA. A seed-and-extend paradigm is applied with the rest of the read being allowed either a specified number of mismatches or a specified base quality score threshold to allow for mismatches. Examples of hash table-based mappers include MAQ (Li et al., 2008a), SOAP (Li et al., 2008b) and BLAST (Altschul et al., 1990, Altschul et al., 1997). Another common indexing method is the FM-index based on a Burrows-Wheeler Transformation (BWT). The BWT was originally a method for compressing data, but can be used to greatly reduce the search space and find the location of a sequence using a prefix-tree. Similarly to the hash-table based tools, BWT-based tools often use a seed-and-extend paradigm using mismatch counts or quality score thresholds. Examples of BWT-based aligners include Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009) and SOAP2 (Li et al., 2009c).

Mapping tools face challenges from constantly evolving sequencing technology. Mapping algorithms must be able to cope with inherent biases in the sequencing platforms and increasing read lengths. As a result there are a large number of mapping tools available, with frequent releases of new tools

designed to keep up with the changing landscape. Pressures from the evolving sequencing technology, the wide variety of downstream applications and constraints on runtime and memory footprints require tools to offer a variety of modifiable parameters to widen the tools' applications. While flexibility allows the experienced user to adapt tools to specific purposes, for the novice user the additional parameters can add to the confusion of which tool to use and how. Multiple mapping tools run with default settings on the same data set invariably return different results, owing to the different methods and default settings used in each tool (Hatem et al., 2013, Bao et al., 2011, Li and Durbin, 2009, Cornish and Guda, 2015). Even when tools are run with parameters set to match one another they can return quite different results in different runtimes with different memory footprints (Hatem et al., 2013). This makes the choice of mapping tool an important first step in a pipeline and one should consider the strengths and weaknesses of each tool and modify parameters to suit the experiment's purpose and resources.

Some tools such as BWA-mem (Li, 2013), GraphMap (Sović et al., 2016), BLASR (Chaisson and Tesler, 2012), Minimap2 (Li, 2018) and NGMLR (Sedlazeck et al., 2018) are able to handle the longer reads and lower accuracy of the latest technologies. BWA-mem is not specifically designed for long reads but for short reads and assembly contigs, however with recommended modifications to the parameters it performs well when mapping long, error-prone reads. BWA-mem works using BWT and a seed-and-extend paradigm as described previously (Li, 2013). GraphMap was designed with long reads in mind, but is designed to be agnostic of error

profiles, it uses multiple gapped seeds to reduce the search space and find a rough region the read should align to, the loci of the seeds are then refined and these are used as anchors which are chained together by the longest common subsequence before the location of the entire read is determined by linear regression. There can be multiple candidate locations per read and the best one for each read is selected at the end of the process (Sović et al., 2016). BLASR has been designed specifically for long reads from single molecule sequencing. Similarly to GraphMap, BLASR generates multiple short matches to give an approximate location of the read, however BLASR does not use gapped seeds and instead relies on seeds with exact matches. Sparse dynamic programming is used to generate a rough alignment from the exact matches and this is then refined (Chaisson and Tesler, 2012). Minimap2 also finds a series of exact matches from a hash table to use as anchors and identifies collinear anchors as chains to find the approximate location of the read and then uses dynamic programming to extend the chains and join the gaps between anchors for base-level alignment (Li, 2018). NGMLR is designed specifically to be able to align across structural variations between long reads and a reference. It splits the reads into 256bp segments, aligns them independently to the reference and groups collinear segments from each read into long segments which are aligned to base-level with dynamic programming with a convex gap-cost scoring scheme. Finally, the highest scoring, non-overlapping alignment for each read is selected (Sedlazeck et al., 2018).

Some tools such as Minimap (Li, 2016) and MUMmer (Kurtz et al., 2004), allow for low accuracy and high speed alignment. While these aren't useful for identifying small variants, they can allow for a very quick, rough alignment between two long sequences which can be visualised. These alignments can be useful for quickly aligning whole genomes to whole genomes, effectively revealing large structural variation between the two sequences. They can also be used to quickly find the genomic context of a long sequence when high levels of accuracy are unnecessary. Both of these tools can handle noisy long-read data.

## 1.8 Variant Calling Algorithms

Most algorithms for variant callers are based on Bayesian probability, they calculate the probability of a genotype given the background data. The main categories of variant caller are: germline callers – used in identification of mutations causing disease or phenotype changes; somatic callers – used in cancer studies which compare tumour cells' sequences with a normal cell's sequence; and structural variant callers – used to identify structural variants (SVs) such as large indels, inversions and translocations.

Callers vary in how they handle data, the metrics they consider and the criteria they use to identify variants (Yu and Sun, 2013). This results in different callers giving different results for the same dataset, making it important to understand some of the major differences between them. Currently, the most popular variant caller is Genome Analysis Toolkit (GATK; McKenna et al., 2010) which is frequently found to be the most sensitive and

specific in comparisons of the performance of different tools and analysis pipelines (Yu and Sun, 2013, Pabinger et al., 2014, Yi et al., 2014, O'Rawe et al., 2013, Warden et al., 2014). Other tools include Atlas2 (Evani et al., 2012), Samtools (Li, 2011) SOAPsnp (Li et al., 2009b), VarScan 2 (Koboldt et al., 2012), SNVer (Wei et al., 2011), Platypus (Rimmer et al., 2014) and FreeBayes (Garrison and Marth, 2012).

Some of the best variant callers have shifted the focus from the genotype to the haplotype. Bayesian formulae are applied to the haplotype to determine which alleles are most likely to be true, this haplotype is then broken down into its constituent variants. Examples of haplotype-based callers are GATK's HaplotypeCaller (GATK-HC; Poplin et al., 2017), FreeBayes (Garrison and Marth, 2012) and Platypus (Rimmer et al., 2014). In Platypus and GATK-HC, local reassembly of haplotype windows using de Bruijn-like graphs allows the caller to be more robust to indels and other structural variants. This also allows for the acceptance of lower accuracy in the mapping stage as the caller will reconstruct the region regardless of the quality of the original mapping. Additionally, Platypus allows for the calling of multiple nucleotide variants (MNVs), SNVs that are consistently found on a single haplotype. Calling MNVs in protein-coding regions will allow for more accurate annotation of the effect these variants have on protein structure and function. GATK-HC best practices involve many stages that rely on the availability of databases of additional resources such as known SNVs and indels to improve the accuracy of the calls. While these resources are available for

humans, many stages of the best practices need to be modified or excluded for other species where fewer resources are available.

Haplotype-based callers have not yet been thoroughly tested in tool comparison studies, though one study found that GATK-HC called similar numbers of SNVs to GATK-UG, but significantly more indels, it is unclear how many of these were false positives (Warden et al., 2014). In a study by Lescai et al. (2014) only 55.9% of GATK-HC indel calls assessed were successfully validated. Hwang et al. (2015) attempted to use a gold standard set of variants to determine which caller was the most accurate on Illumina data, while GATK-HC came out on top with Freebayes and Samtools producing the most false-positives, the gold standard indel set was generated using this GATK-HC and is likely to have biased the results. Cornish and Guda (2015) compared 5 variant callers: GATK-HC, GATK-UG, Samtools, Freebayes and SNPSVM. They found that Freebayes produced the most raw variant calls, but also the most false-positives. GATK-UG was found to be the most sensitive caller, outperforming GATK-HC. As was found by Warden et al. (2014), similar numbers of SNVs were observed, but GATK-HC called more indels. None of the tested callers performed well for indels. They also found that the choice of aligner had an impact on the results, with burrows-wheeler based aligners (BWA mem, BWA sampe, Bowtie2) behaving similarly to one another, and other aligners (MOSAIK and CUSHAW3) producing different variant calls downstream.

Concordance between SNVs called by variant callers is often low (O'Rawe et al., 2013, Yi et al., 2014, Yu and Sun, 2013), although recently the

concordance may be improving due to updates (Hwang et al., 2015) . It has been suggested that best practice is to use more than one variant caller to ensure all true variation is detected (Pabinger et al., 2014, Yi et al., 2014). This is particularly important when looking for specific disease causing mutations, as any SNV could be relevant (O'Rawe et al., 2013). Some tools such as Consensus Genotyper for Exome Sequencing (CGES), have been developed which combine multiple tools, in this case GATK v2.8, Atlas-SNP2, Samtools and Freebayes. This allows the user to choose whether to accept variants called by all four callers; three or more of the four; two or more; or called by at least one caller and provides detailed quality statistics to aid in this decision (Trubetskoy et al., 2015).

Filtering following the calling process is particularly important to reduce the number of false positives, particularly if multiple callers are to be used (MacArthur et al., 2012) or if there are likely to be inaccuracies in the reference genome.

There are several types of large SV including large insertions, large deletions, interspersed duplications, tandem duplications, inversions, translocations, inverted duplications and unbalanced translocations, and combinations of these (Guan and Sung, 2016). In cancer research, methods that do not rely on the reference genome can be used, comparing normal and tumour samples, for example SMUFIN (Moncunill et al., 2014), but most SV callers do utilise mapping to a reference. These callers generally identify discordant mapping of paired reads, i.e. reads that map in unexpected orientation, with unexpected insert size, with ends on different chromosomes,

with one end not mapped, or soft clipped reads. Examples of tools for SV calling include LUMPY (Layer et al., 2014), SVachra (Hampton et al., 2017), Ulysses (Gillet-Markowska et al., 2015), SVMiner (Hayes et al., 2012), GASV (Sindi et al., 2009), CLEVER (Marschall et al., 2012), ClipCrop (Suzuki et al., 2011) and PRISM (Jiang et al., 2012). Mapping algorithms are designed to map concordant reads and mapping across break points for SV detection is often unreliable, impacting the reliability of downstream analyses (Lim et al., 2015). An alternate approach is to assemble contigs from reads and call SVs by comparing these to the reference, however this is computationally expensive and comes with the flaws associated with assembling short-read data. Some tools assemble targeted regions based on abundance of soft-clipped reads, but this again relies on the mapping algorithm's ability to properly map them and sufficient coverage for assembly. Examples of tools that use contig assembly to find SVs include TIGRA (Chen et al., 2014), SVMerge (Wong et al., 2010), SOAPindel (Li et al., 2013b) and CREST (Wang et al., 2011b). SVs such as deletions and duplications can be detected through changes in coverage. Large deletions of sequence present in the reference can be detected, assuming accurate mapping and unbiased sequencing, by identifying regions with half the expected coverage or no coverage depending on zygosity. Duplications where there are two or more copies of a single region in the reference genome can be identified through increased coverage and number of copies estimated from the fold increase in coverage, however these may be obscured by pile ups of reads in repetitive or low-complexity regions. In the case of large insertions, the sequence of the

insertion is missing from the reference genome and so these reads are not mapped and are unlikely to be identified. Using coverage in this way is more reliable using WGS than WES, as WES probes create non-uniform coverage (Lelieveld et al., 2015). Tools that use coverage in this way include CNVnator (Abyzov et al., 2011), ReadDepth (Miller et al., 2011), BIC-seq (Xi et al., 2011), Ximmer (Sadedin et al., 2018) and EXOMESNV (Sathirapongsasuti et al., 2011). All of the methods for SV detection discussed so far, with the exception of cancer studies comparing normal and tumour samples, rely on an accurate and contiguous reference genome. Inaccuracies and contig breaks in the reference will show up as SVs and mask true SVs. Unfortunately, species with poor-quality reference genomes are also likely to be the species for which there is insufficient funding for WGS of large cohorts, making it more difficult to filter out these false SNVs.

While long reads have error rates too high for accurate SNP and small indel calling, they allow for identification of large structural variants based on points where long read alignments consistently break. Using long reads for SV detection has the benefit of being more likely to cross break points and potentially span entire SVs. This has many of the benefits of the contig assembly methods of short read SV calling, but is more accurate. Tools for both PacBio reads and ONT reads have been designed for this purpose, they include PBHoney (English et al., 2014), nanoSV (Cretu Stancu et al., 2017), Picky (Gong et al., 2018), and Sniffles (Sedlazeck et al., 2018). These tools rely on the accuracy of the alignment of the long, error-prone reads and

choice of alignment tool will impact on the results. Even when using long reads, breaks and errors in the reference genome will cause false positives.

## 1.9 Variant Annotation Tools

An important step in the process of filtering called variants is variant annotation. Variant annotation attempts to label variants based on their probable effect on biological function. In its simplest form, this involves identifying the consequence of a variant in the protein-coding region on a gene's transcripts and resultant protein products. Protein-coding genes are the most well understood portion of the genome and where reliable gene models, variant databases and transcript data are available it is relatively simple to predict the effect on proteins, though not necessarily the phenotype. Identifying the impact of variants in the non-coding regions is more complicated as the molecular mechanisms of the effect on phenotype from such variants are poorly understood. NGS studies frequently produce thousands of variants and manual annotation is laborious, error prone and impractical. Several different methods are employed to annotate variants computationally (Ritchie and Flicek, 2014), often in conjunction with one another.

The overlap or proximity of a variant to a known functional element may suggest it has a functional effect, particularly if the variant is in the coding sequence or a splice site, in which case the change caused to protein products can be predicted (Adzhubei et al., 2013, Hu et al., 2013, Wang et al., 2010, McLaren et al., 2010). Variants that lie in certain sequence motifs

also have the potential to be disruptive and can be identified by annotators (Wang et al., 2010), these including binding sites for proteins that bind to DNA or RNA, disruption of which may alter regulation of gene expression. The use of multiple sequence aligners which identify whether or not the region has experienced evolutionary constraint may be used to identify potential functional regions. If a sequence has been highly conserved through evolutionary history it suggests that mutations in the region have been selected against and that the region has some important function, suggesting a variant identified in that region is likely to be deleterious (Adzhubei et al., 2013, Adzhubei et al., 2010, Siepel et al., 2005, Kumar et al., 2009). Supervised learning algorithms in which a naïve algorithm is trained on a list of variants that are known to be functional and a list of variants that are known to be benign have been used to annotate variants (Adzhubei et al., 2010, Adzhubei et al., 2013, Schwarz et al., 2010, Ritchie et al., 2014). The algorithm can then analyse novel data and predict whether variants are functional or not. While this method has been used successfully it is often not clear to the user what feature has caused the algorithm to flag a variant as functional. Finally, annotators can use phenotype association information to flag variants that appear to be associated with the phenotype of interest – a particularly useful method when looking for disease related variants (Yandell et al., 2011, Hu et al., 2013, Wu et al., 2011).

Many annotators rely heavily on the availability of databases of variants, transcripts and conserved sequences, both the choice of database and the choice of annotation tool can impact on how the variant is annotated

(McCarthy et al., 2014). Importantly there is no universal definition for the different possible consequence types for variants and many annotators differ in how they classify them. Additionally, while some annotators report all possible consequences (e.g. McLaren et al., 2010, Cingolani et al., 2012), leaving the user to prioritise them, some only report the one that is considered the most severe and these classifications and priorities may differ between tools. This can lead to different annotators assigning different consequences to the same variant even within the coding region (McCarthy et al., 2014). Similarly, if a variant is located in a region where several transcripts overlap, the tools differ in how these are reported. To limit these differences and make the analysis pipeline more accessible to researchers without bioinformatics training, automated tools for annotating variants from exome data have been designed (Liu et al., 2012, Mutarelli et al., 2014) which may be particularly useful for medical professionals.

Most annotators are designed for use on human data, though some are capable of analysing data from other species (Cingolani et al., 2012, McLaren et al., 2010). Notably Ensembl's Variant Effect Predictor (VEP) can be used on any species in the Ensembl genome browser and its output provides links to loci, transcripts and genes in the genome browser (McLaren et al., 2010).

Some annotators, such as Vcfanno (Pedersen et al., 2016), will attempt to annotate SVs, however this is difficult and imperfect as the precise location and nature of SV breakpoints are often not known.

## 1.10 Consequences of Variants on Proteins

During translation from messenger RNA (mRNA) to proteins, nucleotides are read in sets of three, these sets are called codons. There are 64 possible codons combinations of the four nucleotides (figure 1-9) which translate into the 20 amino acids, the start codon (AUG) and the stop codons (UAA, UAG, UGA). If a codon is changed or the reading frame shifts, this can have a number of effects on the resulting protein. Following annotation of small variants (SNPs and indels), the variant is labelled with a consequence and a predicted severity. Annotators will annotate variants across the whole genome and many variants in promoters, enhancers, splice sites, and other regions of the genome may have a phenotypic effect, but the protein-coding regions are the most well understood and can often be more readily associated with a phenotype in genes with known functions. The main consequences of variants in protein coding regions are:

- Synonymous - a SNP that does not change the coded amino acid due to redundancy in codon translations (figure 1-9). Often a substitution of a SNP in the last nucleotide of a codon does not change the amino acid.
- Missense - a SNP that changes one amino acid in the protein.
- Nonsense or stop-gain - a SNP that introduces a premature stop codon, truncating the protein.
- Start-loss - a SNP that alters the start codon and may prevent the protein from being translated.

- Frameshift - an indel that has a length that is not divisible by 3 and causes the reading frame to shift. This causes significant changes in coded amino acids for the remainder of the transcript and usually modifies several amino acids and introduces a stop codon somewhere in the sequence.
- In-frame insertion/deletion - an indel with a length that is a multiple of 3 that may insert or delete one or more amino acids from the protein, but does not affect the rest of the codons.

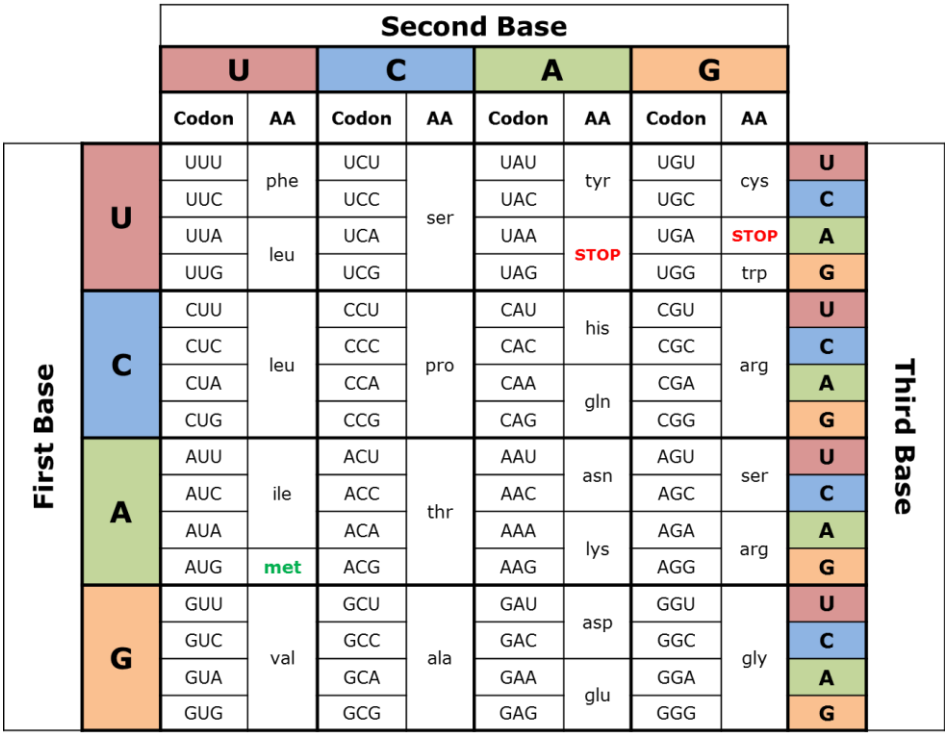



**Figure 1-9- Codon translation table. Note that there is a high level of redundancy for the translation of the third base of the codon in particular.**

Synonymous SNPs are considered the lowest severity variant in protein coding sequence. While there is some evidence that certain synonymous variants can have a strong phenotypic effect, including disease phenotypes, due to alterations in mRNA splicing (e.g. Gallego-Bustos et al., 2016, Claverie-Martin et al., 2015, Agashe et al., 2016), the majority of synonymous variants are thought to be benign or very minor in their impact.

Missense SNPs and in-frame indels are considered moderate impact variants. There are numerous examples of these variants causing phenotypes including disease phenotypes (Yntema et al., 2002, Craig et al., 2009) and non-detrimental phenotypes (Fontanesi et al., 2006), but a phenotype is not always apparent.

The highest impact coding region variants are those which truncate the protein or prevent initiation of translation such as stop-gain, start-loss and frameshift variants. If a protein is truncated early in translation it is likely to be non-functional. Protein truncating variants, sometimes called loss-of-function (LoF) variants, are known to be causative of a number of disease phenotypes (e.g. Cutting, 2014, Aartsma-Rus et al., 2016). Logically, it would be expected that true LoF variants would be rare, particularly in the homozygous state, and largely confined to non-essential genes. Research by MacArthur et al. (2012), however, found that on average healthy humans carry ~100 putative LoF variants with ~20 genes homozygous for putative LoF variants. While some of these variants may cause non-detrimental variation in phenotypes, the variants found in the MacArthur et al. (2012) study were not limited to non-essential genes suggesting potential LoF-tolerance through

redundancy of genes and protein functions in humans. Some variants in the study had been previously associated with disease phenotypes that were not present in the subjects, suggesting there is some unknown genomic or environmental context in which these variants cause disease. Alternatively there may be some mechanism of correcting some of these variants at the mRNA stage and rescuing the protein. Groenen et al. (2012) identified >150 putative LoF variants in 48 pigs representing 8 domestic pig breeds and the wild boar, averaging only 30 per individual. This is fewer than were found in humans, despite the relatively high nucleotide diversity in pigs (Bosse et al., 2012), the authors suggest this reflects the higher effective population size in pigs compared to that of humans. Of course, it may also reflect the limitations of a draft assembly and a higher proportion of both false positives and false negatives when seeking to identify putative LoF variants.

The results of variant annotation should be treated with caution, low severity variants can be causative of disease and high severity variants can have no observable effect on phenotype. However, high severity variants are more likely to cause a detectable phenotype and for the protein coding region these remain the most promising starting point for finding causative variants.

## 1.11 Application of whole exome sequencing

### 1.11.1 WES Applications in Humans

The first successful use of WES to diagnose and alter treatment in a human patient was in the identification of the causal variant of a rare form of Inflammatory Bowel Disease (IBD) in an infant (Worthey et al., 2011). In this

case conventional diagnostics had failed to find an explanation for the patient's severe symptoms and doctors needed to understand the underlying cause of the symptoms before they could decide how to treat the child. A multidisciplinary team combined clinical phenotyping, exome sequencing, bioinformatics and functional studies, and eventually found the causative mutation which influenced the treatment of the child. Through filtering and manual inspection the candidate pool was reduced to 70 genes exhibiting hemi- or homozygous variants. Of these only eight variants were novel and predicted to be damaging to protein function. Analysis of evolutionary conservation identified two variants that were highly conserved and one of these had a high null genotype frequency. This left a single hemizygous, non-synonymous variant in the *X-linked inhibitor of apoptosis (XIAP)* gene. The child was diagnosed with X-linked lymphoproliferative disease 2, exhibiting a novel IBD-like manifestation caused by loss of tolerance to commensal organisms in the digestive system. This allowed for effective treatment through allogeneic hematopoietic progenitor cell transplant. The unusual manifestation of the condition meant that a diagnosis for the patient was unlikely without the exome sequencing data.

Following the success of this initial diagnosis, exome sequencing has been used extensively to diagnose novel diseases and find novel causative mutations of known diseases. Exome sequencing is useful in human medicine for diagnosis of particularly difficult to diagnose patients, diagnosis of young patients who may not exhibit a full spectrum of symptoms yet (Iglesias et al., 2014), prenatal diagnosis (Xu et al., 2014, Iglesias et al.,

2014) and early diagnosis of debilitating disease (Sassi et al., 2014, Bras and Singleton, 2011). In addition to finding a diagnosis, finding the causative mutation can allow for alteration of treatment, prevention of further invasive testing, accurate prognoses, and confirmed diagnoses which are essential for eligibility for benefits and access to clinical trials (Taneri et al., 2012, Rabbani et al., 2012, Iglesias et al., 2014, Grossmann et al., 2011) and in the future may allow for more targeted treatment. Recently, a study found that of 278 infants in intensive care units that were referred for exome sequencing, 36.7% were given a molecular diagnosis and over half of those diagnosed saw informed redirection of care, initiation of new subspecialist care, medication/dietary modifications, and furthering life-saving procedures (Meng et al., 2017a). The study also found that for roughly half of infants who were deceased, genetic disorders were diagnosed allowing risk of recurrence counselling for the parents.

Exome sequencing in human medicine benefits from the availability of large databases of known SNPs, known pathogenic variants and control genomes, during analysis variants found in these databases can generally be excluded when looking for novel variants significantly reducing the variant pool. The Exome Aggregation Consortium (ExAC) has created a user-friendly database containing the exome sequences of over 60,000 unrelated individuals, which is freely available from Exome Aggregation Consortium (ExAC) (2014). The exomes have been analysed using a uniform bioinformatic pipeline and are from individuals with adult-onset diseases. This provides researchers with a large set of reference exomes which should be free from homozygous

variants that cause childhood-onset Mendelian diseases. The database provides a wealth of information such as depth of coverage, genotype quality, allele frequency and variant consequences. The filtering capabilities and large number of exomes will simplify the process of prioritising variants when using exome sequencing as a diagnostic tool, particularly in children.

Where many unrelated, affected individuals are available, an 'overlap' strategy can be used to simplify analysis and search for common variants likely to affect gene function (Johansson et al., 2012). If there is a known inheritance pattern this can be used to search for specific genotype zygostiy. Sequencing of multiple affected related individuals can also increase the power of the analysis (Johnson et al., 2010, Shi et al., 2011). In rare diseases, case-parent trios can be used to exclude non-pathogenic variants found in the parents (Smith et al., 2014). However, single patient sequencing may be sufficient to identify the causative mutation (Worthey et al., 2011, Xu et al., 2014, Wang et al., 2011a). Analysis examining how conserved an amino acid sequence is through evolution and between genes in a family can help to increase the confidence of a found mutation having a deleterious effect (Xu et al., 2014, Wang et al., 2011a, Johansson et al., 2012, Sassi et al., 2014, Smith et al., 2014).

Examples of diseases for which exome sequencing has been used to detect a causative variant include Leber Congenital Amaurosis (Wang et al., 2011a), Alzheimer's Disease (Sassi et al., 2014), Maturity-Onset Diabetes of the Young (Johansson et al., 2012), High Myopia (Shi et al., 2011), Autosomal Recessive Polycystic Kidney Disease (Xu et al., 2014),

Amyotrophic Lateral Sclerosis (Johnson et al., 2010), immunodeficiency leading to infection with human herpes virus 8 causing Kaposi Sarcoma (Byun et al., 2010), Acromelic Frontonasal Dystois (Smith et al., 2014) and a number of cancer predisposition mutations (e.g. Kiiski et al., 2014, Snape et al., 2012, Greif et al., 2012, Yan et al., 2011, Cai et al., 2015) and variants linked to developmental disorders (Wright et al., 2015, Gecz and Corbett, 2015).

### 1.11.2 WES Applications in Non-Human Species

As with exome sequencing in humans, exome sequencing in other mammals has been largely aimed at discovering variants associated with health traits. WES has been used in conjunction with a genome wide association study to identify a frameshift mutation causing blindness in Phalène dogs (Ahonen et al., 2013). In cattle (*Bos taurus*), WES has successfully been used to identify strong candidate variants for haplotypes relating to reduced fertility rates in Holsteins which can be used to selectively breed against these detrimental haplotypes (McClure et al., 2014). The kit used in this study can also be applied to other bovid species. Cosart et al. (2011) demonstrated that the kit could be successfully used to capture the exomes of, and identify SNPs in, zebu (*Bos indicus*) and American Bison (*Bison bison*). This transferability of exome capture kits, as also demonstrated in studies involving the sequencing of Neanderthals and non-human primates using human capture kits (Burbano et al., 2010, Vallender, 2011), is possible because despite millions of years of divergence, functional elements tend to be highly conserved.

Plant genomes can be extremely complex, repetitive and are often polyploid, as a result high quality reference genomes are often not available and even where they are, the size and amount of repetitive content makes them expensive to sequence for variant discovery. Bread Wheat (*Triticum aestivum*) has an allohexaploid (AABBDD) genome around 17 Gb in size. While the wheat genome has been assembled (Clavijo et al., 2017), it is likely too large for subsequent resequencing studies at the current price of sequencing and data storage. Wheat is an extremely important crop for both human and livestock consumption and genetic improvement is slow. An exome capture kit has been designed based on the accumulated transcriptome data for wheat (Winfield et al., 2012). The capture region for this kit is 56.5Mb, which is around the lower estimated size of one diploid wheat exome, and owing to similarity between the three genomes this may be sufficient to capture most of the exome data from the whole allohexaploid genome. The kit has allowed for discovery of previously unidentified markers in the genome which can be used in future genetic studies and marker assisted selection (Allen et al., 2013). The kit has also been used to identify induced mutations in the genome to aid in studies investigating gene function, a use that was also applied to the rice (*Oryza sativa*) exome in the same study (Henry et al., 2014) and the soybean (*Glycine max*) exome in a separate study (Bolon et al., 2011). WES in soybean has also been used to identify unwanted intracultivar genetic heterogeneity in the exome that may affect the plant's phenotype (Haun et al., 2011).

The genome of barley (*Hordeum vulgare* L.) has recently been assembled (Mascher et al., 2017). Barley's genome is smaller than wheat's at around 5Gb, but is still larger than is practical and contains many repetitive elements. Previously, a gene space assembly was produced (The International Barley Genome Sequencing Consortium, 2012) and a barley exome capture kit has been developed based on this assembly (Mascher et al., 2013). The kit has since been used to identify a mutation involved in early maturation, a trait relevant to production (Pankin et al., 2014). The kit has also been used to differentiate between markers of *H. vulgare* L. and *H. bulbosum* L., *H. bulbosum* L. is a wild species that has superior pathogen resistance and tolerance compared to the domestic species and the two can be crossed to improve the domesticated crop, however negative linkage drag on production traits has hampered its use in elite barley lines. Using exome sequencing to identify specific markers can allow selective crossing to be used to incorporate the beneficial variants without incorporating linked variants that are detrimental to production (Wendler et al., 2014).

Exome capture in barley has also been used to identify a gene causative of many-noded dwarfism (mnd) using mapping-by-sequencing (Mascher et al., 2014). The mnd phenotype is a shorter plant with more, narrower leaves than the wild type. The mutant in this study was created using X-ray mutagenesis, a technique which often causes large deletions. An $F_2$ population between mutant and wild type phenotypes was created and 18 mutant individuals and 30 wild type individuals were exome sequenced. From these sequences SNPs were identified and allele frequencies of these were used to identify an

allele over represented in the mutant group. Researchers queried the sequencing reads for exome targets that were present in the wild type but not the mutant. This lead to the identification of a candidate gene (MLOC_64838.2, now HvMND), which has a homologue known to play a role in a similar phenotype in rice. Screening of other mutants showing this phenotype found a variety of null mutations in this gene. The family to which this gene belongs is known to have effects on important production traits and may include good selection targets to improve production.

In addition to plants important in food production, black cottonwood (*Populus trichocarpa*) has had an exome capture kit designed (Zhou and Holliday, 2012). *P. trichocarpa* is a model organism and was the first tree to have its whole genome sequenced (Tuskan et al., 2006). The tree is used in lumber production and in cosmetics. There is potential to use the identified SNPs to improve production in this species.

So far, exome sequencing has not been widely used to identify variants related to production in mammals. However, Robert et al. (2014) have designed exome capture probes for the pig, used these to sequence the exomes of 96 healthy pigs and using bioinformatic tools identified potentially deleterious variants in these sequences. Bioinformatic analysis identified 236,608 high confidence predicted variants and 28,115 predicted indels in the target region. This work revealed notable gaps in the current Ensembl *S.scrofa* genome annotation and identified a large number of potential protein truncating variants. As the pigs tested were healthy, it is possible that some of these protein truncating variants are having a phenotypic effect on traits

other than those relating to the health of the pigs, such as those relevant to production. This work is an important step in identifying phenotype altering variants in the pig: a production animal and medical model.

### 1.11.3 Benefits of WES over WGS

With the price of sequencing falling as rapidly as it has done over the past decade (figure 1-10), questions have been raised concerning WES's usefulness in the era of affordable WGS. The costs of WES consist of the cost of the capture plus the cost of sequencing, whereas WGS consists only of the sequencing costs. If we assume that the cost of capture remains fixed, then as the costs of sequencing fall, the cost of WGS will approach the cost of WES. However, at present that is not the case and it is unlikely that the cost of sequence capture will not also reduce. Illumina's HiSeq X platform offers a cost-per-Gb far less than other platforms. However, even given that advantage, the $1000 price tag for a 30X human genome is two to three times the cost of a 40X human exome (depending on scale). While WGS does have benefits over WES, the cost of this technology is more than simply the price of sequencing. Sequencing technology has been improving at a much faster rate than would be predicted by Moore's law (a prediction of improvement in computing hardware, but often also applied to other technologies), but the technology for storing and analysing the data has not seen a matching acceleration in improvement (Mardis, 2010, Sboner et al., 2011). WGS produces around one hundred times the data that WES does at the same coverage. The infrastructure needed to store, manage and analyse data significantly increases the costs of WGS.

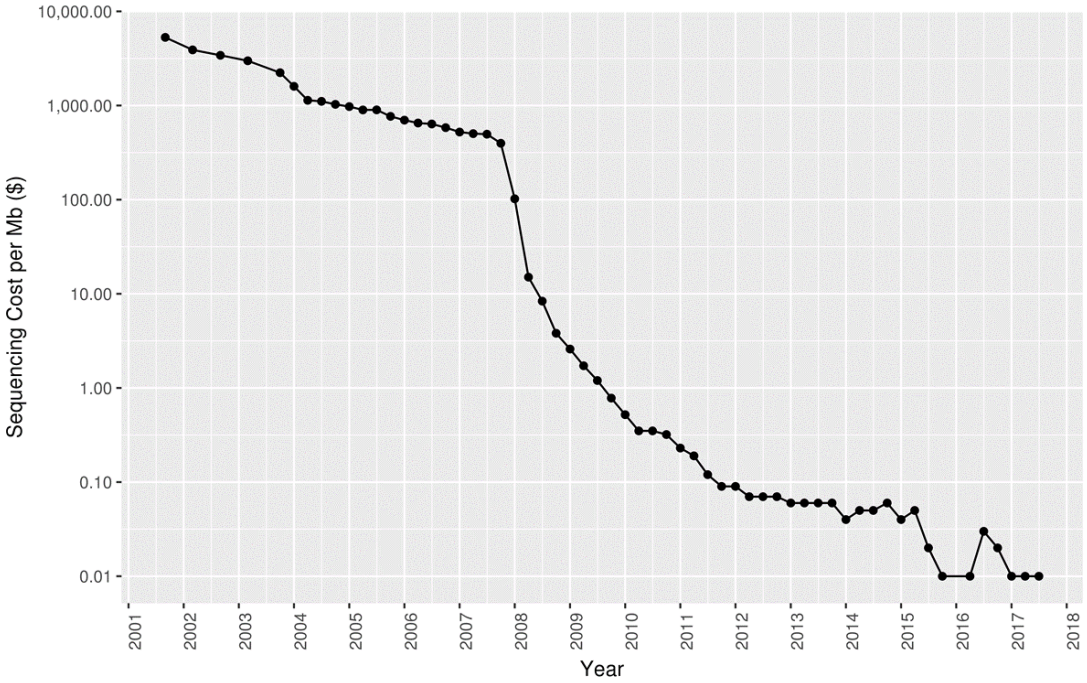



**Figure 1-10- The falling cost of sequencing. Data from the NHGRI Genome Sequencing Program (Wetterstrand, 2017)**

WGS produces a much larger number of variants than WES does, not only because of the size of the sequencing space, but because regions outside the exome are less well conserved. While this number might include a variant of interest, the larger data set significantly increases the computational burden for analysis. Additionally variation in non-coding regions is currently more poorly understood than variation in the coding region (Ward and Kellis, 2012, Mu et al., 2011, Maurano et al., 2012), making it more difficult to predict which might be relevant to a trait of interest in WGS datasets. The majority of causative variants identified so far in Mendelian disease have been found in coding regions (Botstein and Risch, 2003), although ascertainment bias is likely to play a role in this. Variants in regulatory regions are increasingly being associated with quantitative traits (Schaub et al., 2012) including variants linked to economically important traits such as muscle growth (Laere et al., 2003, Cockett et al., 2005). With the continuous decrease in sequencing cost, new studies making use of WGS to investigate causative variants will lead to the discovery of additional mutations in regulatory elements that contribute to the pool of disease-associated variants. In that context, the sampling bias currently observed towards coding variants is likely to be reduced by WGS investigations of non-coding genomic regions.

Cost is not the only consideration. WGS covers the whole genome at more consistent coverage than WES, can provide more accurate detection of structural variants and does not exhibit reference sequence bias caused by probe sequences in WES (Majewski et al., 2011, Meynert et al., 2014,

Belkadi et al., 2015). Recent studies have highlighted the importance of promoters (The Fantom Consortium et al., 2014) and enhancers (Andersson et al., 2014) in a range of different cell types, and these are not traditionally captured by exome sequencing.

WGS is now being widely applied in human studies, including the sequencing of large cohorts (Erikson et al., 2016, Gilly et al., 2018), and plans are in place to sequence the genomes of 500,000 individuals in the UK Biobank (Baras, 2018). WGS will eventually take a leading role in genome interrogation in agriculture, however to sequence large cohorts accurately and at high coverage, it will likely have to wait for costs to reduce and data storage and analyses to improve before it can be feasibly used to its full potential. In the meantime, WES provides many of the benefits of WGS with lower storage requirements and computational burden, at an affordable price. This is particularly useful in large scale studies, for example, a recent study sequenced the exomes of over 9,000 people (Schick et al., 2015). It is also useful for the sharing of information such as in the ExAC database where there are over 60,000 human exomes stored. WES will likely also remain the method of choice in species with exceptionally large genomes, for example in some polyploid plant species.

An alternative to carrying out high-coverage sequencing on a large cohort of animals is to use genotyping-by-sequencing to target specific markers in the genome, which can give comparable accuracies to SNP array at as low as ~1X coverage (Gorjanc et al., 2015). This is a much cheaper alternative to WGS thanks to the potential to multiplex high numbers of individuals. This

can be further extended to low-coverage WGS of large numbers of individuals with high-depth sequencing of minimal key individuals which maximally represent the genetic diversity of the population (Gonen et al., 2017, Ros-Freixedes et al., 2017). The high coverage sequencing allows for the most common haplotypes in the population to be assembled, and the accumulated low-coverage sequencing can capture the less common haplotypes and variants allowing for accurate imputation of whole genomes to the population (Ros-Freixedes et al., 2017). As it is already common practice for SNP array data to be collected and pedigrees to be recorded in agricultural species, this can be used to select key individuals that share large numbers of haplotypes with the rest of the population for deep sequencing and allows for allocating the highest proportion of a fixed sequencing budget to the most informative individuals (Gonen et al., 2017). While an interesting option, this low coverage sequencing approach is still theoretical and has not yet been applied to real data.

## 1.12 Conclusions

Genome sequencing and assembly have come a long way since the first human genome, and with reducing costs and advances in technology these projects are no longer limited to large consortia, even small labs can afford to sequence their species of interest. While assembly tools have improved greatly, it has been necessary to constantly adapt these tools to work with the different biases and characteristics of the data produced by new and improving sequencing technologies. Assembly of large genomes, while

cheaper, faster and easier than it once was, can still not be considered trivial. However, as these tools and technologies continue to improve it is essential that scientists are not only able to sequence their species of interest, but to assemble its genome well. Reference genomes greatly improve the chances of finding genetic variants that influence a phenotype, but inaccuracies in references lead to false-positives in downstream analyses. The introduction of errors in such an early stage of analysis greatly increases the amount of work needed to find meaningful results in downstream analyses and the extreme filtering that must be used increases the chances that a true-positive, important variant will be removed from the data set.

Genomic analysis is computationally expensive and large scale studies using whole genome sequencing are less feasible for large cohorts at the current cost of sequencing. Exome sequencing is a technology that allows interrogation of the most well-understood portion of the genome: the protein-coding region and functional elements. Variants of interest don't necessarily fall within the exome, but most of the known variants responsible for Mendelian disease have been found in the coding region and the target region can be extended to include other regions of interest. While the falling price of sequencing may soon make WGS the more attractive of the two techniques, data handling and downstream analysis increases the cost.

The amount of data produced by WES is far more manageable than WGS, particularly for small research groups and groups studying organisms with large genomes. WES has established itself as an important method in disease gene identification in humans, and increasingly in domestic species.

The applications of WES in crop research is allowing genomic techniques to be used in species with complex genomes, potentially identifying variants important for production that can be incorporated into marker-assisted selection. At present, WES is a useful and powerful method for variant discovery within coding regions offering most of the benefits of WGS while allowing for easier and faster analysis of the data produced.

The usefulness of both WES and WGS depend on the quality of the reference assembly that the sequenced reads are mapped to. In the future, the improvement of reference assemblies, bioinformatic tools and sequencing technology will be necessary to improve the power of variant discovery techniques.

Pigs are an important species for both the food industry and for medical modelling. Pork is the most consumed meat globally and demand will likely increase as the population grows and wealth in developing countries increases. Pigs are far more similar biologically to humans than the other commonly used animal models. If null mutations are found in the breeding population that through purging could improve pork production or through selection could cause a phenotype that mimics human disease, these can be used to breed superior producers or medical models.

## 1.13 AIMS AND OBJECTIVES

This project aims to find genetic variants that may be of relevance to the pork production industry or be of use in medical modelling and to provide improved resources for genomic analysis in the pig. In order to do this, the following objectives will be addressed:

1. An assessment of the quality of the pig reference genome, Sscrofa10.2 will first be carried out to identify any regions that may be poorly assembled and likely to increase the chances of false-positives in analyses.
2. Following identification of these regions, they will be applied as a filter to exome sequencing data previously produced by Robert et al. (2014). These data from the exomes of 96 individuals contain over 236,000 SNPs and 28,000 indels, many of which will be benign or false positives. Filtering is needed to identify potentially phenotype-altering variants. In addition further filtering based on variant annotation and phenotype association will be carried out to identify putative detrimental variants.
3. An alternative approach to avoiding false-positives will be used to identify putative causative variants through sequencing the genomes of a trio of individuals with filtering based primarily on the genotypes of the three individuals and the predicted consequences of the identified variants.

4. Finally, in order to reduce the filtering required in similar analyses in the future, a new and more accurate reference genome will be produced using long-read sequencing technology.

# CHAPTER 2: IDENTIFICATION OF LOW-CONFIDENCE REGIONS IN THE PIG REFERENCE GENOME (SSCROFA10.2)

---

*"Never mind what should be or what might be or what ought to be. It's what things are that's important."*
- Terry Pratchett, Wyrd Sisters

## 2.1 Introduction

Identifying natural sequence variants and testing the variants for associations with phenotypic variation is key to forward genetics approaches to linking genotype to phenotype and subsequent predictions in genomic selection. Effective variant discovery, however, is dependent upon the quality of the re-sequence data and the framework, typically the reference genome for the species of interest, against which the variants are discovered. Therefore in order to identify high quality, usable variant information among variant calling data, they must be filtered. It is common to filter variants based on alignment quality scores and sequencing depth to remove any false positives caused by sequencing or alignment errors. In research on rare human diseases or phenotypes, openly available variant databases mean that further filtering can be done to remove alleles commonly observed in healthy individuals, greatly reducing the number of candidate causal variants. For other species, however, hard filtering must be used to cut down on candidates and remove false-positives. Hard filtering involves stricter thresholds on a number of filters. Filters recommended by GATK include QualityByDepth (variant confidence divided by the unfiltered depth of non-hom-ref samples), FisherStrand (This is the Phred-scaled probability that there is strand bias at the site), StrandOddsRatio (estimate strand bias using a test similar to the symmetric odds ratio test), RMSMappingQuality (the root mean square mapping quality over all the reads at the site), MappingQualityRankSumTest (the u-based z-approximation from the Rank Sum Test for mapping qualities) and ReadPosRankSumTest (the u-based z-approximation from the Rank

Sum Test for site position within reads). How strict these filter thresholds need to be is generally decided by the researcher. For many non-human species the available reference genome sequence assembly is only of draft quality and is more likely to contain errors and missing sequence. This limitation of the reference genome increases the likelihood of false positives making these filters particularly important.

Following exome sequencing on a cohort of 96 pigs, and hard filtering recommendations for exome sequencing data, Robert et al. (2014) were left with a large number of putative variants including >236,000 SNPs and >28,000 indels. In order to find any variants that might be of importance to the pig breeding industry or for medical modelling, these needed to be filtered further and prioritised. Both users of the draft pig genome sequence (Sscrofa10.2; Groenen et al., 2012) and leading members of the Swine Genome Sequencing Consortium under whose auspices the draft genome assembly was produced have identified and recognised flaws in the Sscrofa10.2 assembly (A. L. Archibald, personal communication). Therefore many of the putative variants discovered in the exome data and in whole genome shotgun sequence data, including variants deposited in the public variant databases (dbSNP, European Variation Archive (EVA)), may be unreliable. If regions in the genome can be identified as low-confidence or poor mappability regions, any of the exome variants that fall in these regions can be filtered out to reduce false-positives.

The following publication was written by the author of this thesis with suggestions and minor revisions from co-authors and reviewers. All methods

were carried out by the author with guidance from supervisors. It is published in Frontiers in Livestock Genomics.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 November 2015
doi: 10.3389/fgene.2015.00338

# Identification of Low-Confidence Regions in the Pig Reference Genome (Sscrofa10.2)

*Amanda Warr[1], Christelle Robert[1], David Hume[1], Alan L. Archibald[1], Nader Deeb[2] and Mick Watson[1]**

[1] Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK, [2] Genus plc., Hendersonville, TN, USA

Many applications of high throughput sequencing rely on the availability of an accurate reference genome. Variant calling often produces large data sets that cannot be realistically validated and which may contain large numbers of false-positives. Errors in the reference assembly increase the number of false-positives. While resources are available to aid in the filtering of variants from human data, for other species these do not yet exist and strict filtering techniques must be employed which are more likely to exclude true-positives. This work assesses the accuracy of the pig reference genome (Sscrofa10.2) using whole genome sequencing reads from the Duroc sow whose genome the assembly was based on. Indicators of structural variation including high regional coverage, unexpected insert sizes, improper pairing and homozygous variants were used to identify low quality (LQ) regions of the assembly. Low coverage (LC) regions were also identified and analyzed separately. The LQ regions covered 13.85% of the genome, the LC regions covered 26.6% of the genome and combined (LQLC) they covered 33.07% of the genome. Over half of dbSNP variants were located in the LQLC regions. Of copy number variable regions identified in a previous study, 86.3% were located in the LQLC regions. The regions were also enriched for gene predictions from RNA-seq data with 42.98% falling in the LQLC regions. Excluding variants in the LQ, LC, or LQLC from future analyses will help reduce the number of false-positive variant calls. Researchers using WGS data should be aware that the current pig reference genome does not give an accurate representation of the copy number of alleles in the original Duroc sow's genome.

**Keywords: missassembly, copy number variable regions, structural variation, draft assemblies, false positives**

OPEN ACCESS

***Edited by:***
*Peter Dovc,*
*University of Ljubljana, Slovenia*

***Reviewed by:***
*Simone Eliza Facioni Guimaraes,*
*Universidade Federal de Viçosa, Brazil*
*Martien Groenen,*
*Wageningen University, Netherlands*

****Correspondence:***
*Mick Watson*
*mick.watson@roslin.ed.ac.uk*

***Specialty section:***
*This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics*

***Received:*** *11 August 2015*
***Accepted:*** *12 November 2015*
***Published:*** *27 November 2015*

***Citation:***
*Warr A, Robert C, Hume D, Archibald AL, Deeb N and Watson M (2015) Identification of Low-Confidence Regions in the Pig Reference Genome (Sscrofa10.2). Front. Genet. 6:338. doi: 10.3389/fgene.2015.00338*

## INTRODUCTION

Contemporary genetics research benefits from genomics tools and resources, including DNA sequencing and single nucleotide polymorphism (SNP) chips, which facilitate detailed quantitative molecular characterization of genetic variation at the population and individual level. A high quality reference genome sequence for the species of interest is an invaluable asset for the discovery of molecular genetic variants. Most reference genome sequences for species with large, complex genomes are incomplete representations of the genome sequence of a single individual or a small number of individuals. Given the extent of insertion/deletion (indel) polymorphisms and copy

number variation (CNV) within species, such individual reference genomes do not contain all the sequences present in the species of interest. Thus, there are two major flaws in the current single linear model for reference genomes as a framework for discovery and analysis of genetic variation: (1) errors and gaps in the reference genome assemblies most of which are incomplete drafts; and (2) using a haploid genome of one individual to represent the genome(s) of a species. In this paper we focus solely on the former.

Studies that employ variant calling from sequencing data to find variation in the genome produce large variant call sets (Robert et al., 2014; Belkadi et al., 2015; Bianco et al., 2015; Gudbjartsson et al., 2015). Most of these calls will be either false-positive, not relevant to the phenotype under investigation or benign (MacArthur et al., 2012). Failure to detect true variants (i.e., false-negatives) will also occur either as a result of insufficient sequence depth or gaps in the reference genome (real or technical). Filtering these datasets reduces the number of variants to a level which can be validated, however, in the process researchers risk discarding the variants they are looking for.

Many applications of high throughput sequencing rely heavily on the accuracy of the available reference genome for the species. Errors in the reference genome increase the number of false-positive variant calls in data, resulting in a need for more stringent filters which may increase the risk of removing true-positives. Shortcomings in the reference genome will also increase the risk of missing true variants (i.e., false-negatives). The human genome is more accurate than that of many other species and more resources are available to aid in the filtering of false-positive variants. Many reference genomes have a draft status and gaps and misassemblies are not uncommon (Kelley and Salzberg, 2010). Identifying misassembled regions in the reference genomes of non-human species and excluding them from analysis will help to reduce false-positives in variant calling data.

Whole genome sequencing (WGS) produces fairly consistent coverage across the genome (Belkadi et al., 2015), however, the PCR step in the Illumina library preparation pipeline is known to introduce bias, particularly in regions of high or low GC content (Kozarewa et al., 2009). Modifications have been introduced to protocols to reduce this bias (Aird et al., 2011), however, sequencing depth and quality in GC-rich and -poor regions remain unreliable when using protocols involving a PCR step. Previous work has shown that CNV can be accurately detected in WGS data by looking for areas of excessively high or low read counts following adjustment for GC content (Yoon et al., 2009; Zhang and Backstrom, 2014). To identify misassemblies in the chicken genome, a previous study used a pool of multiple birds to account for true variation between individuals, treating regions where all individuals show low read counts as false tandem duplications (Zhang and Backstrom, 2014).

In this paper, we look to identify low-confidence regions in the reference genome assembly Sscrofa10.2 using WGS reads from T. J. Tabasco (Duroc 2-14), the Duroc sow whose DNA was used in the assembly (Groenen et al., 2012). The assembly was constructed using a BAC-by-BAC method, covers 18 autosomes and 2 allosomes (with the Y chromosome constructed separately from the DNA of male pigs), and contains many gaps and sequences on unplaced scaffolds. Ideally, an individual's sequencing reads mapped to that individual's own assembled genome would show no true structural variation and any areas of structural variation could be considered a misassembly. But the reference genome is a haploid representation and cannot reflect areas of true heterozygous structural variation accurately. However, a conservative approach would treat variant calls in these areas as low-confidence until further verified. Regions with no structural variation between the sequencing reads and the reference genome can be considered high-confidence.

In addition to using coverage to detect potential duplications or collapses, we use other indicators to identify different kinds of structural variation such as inversions, deletions and insertions as has been done previously to identify potentially disease causing structural variation in human genomes (Tuzun et al., 2005). Illumina paired-end sequencing generates read-pairs from the same DNA fragment that are a known distance apart (usually following a normal distribution), and in a known orientation with respect to the reference genome. Therefore, when read pairs are mapped to a reference, if they are not in the expected orientation, or are an abnormal distance apart, this may also be an indication of errors in the assembly.

Finally, when mapping reads from the same animal to the reference genome created from that animal, there should be no homozygous variant calls.

In this work, regions with abnormally high or low coverage (LC), with high proportions of reads with unexpected insert sizes or a high proportion of reads which were improperly paired were identified. In addition, SNP and indel calling was carried out. Regions were considered low quality (LQ) if they had high coverage, a high proportion of unexpected insert sizes or improperly paired reads or if they were in proximity to a homozygous variant. LQ regions are the most likely to represent misassemblies in the genome. Regions which had LC were analyzed separately; these regions may not necessarily be misassembled, but have poor coverage and may therefore be unreliable for accurate variant calling. Both regions were also analyzed together in a combined dataset (LQLC).

Following identification of regions of the reference which may be unreliable, publicly available data sets were downloaded and overlap with the regions calculated. The data sets downloaded were the coding region, dbSNP variants, copy number variable regions (CNVRs) identified by Paudel et al. (2013) using a method that assesses read depth, and gene predictions based on data obtained using RNA-seq methods. These data sets allowed for identification of the proportion of the coding region overlapping the unreliable regions, and to assess how commonly used methods of SNP and indel calling, CNVR calling and RNA-seq may have been affected by unreliable regions of the genome assembly. We would expect the coding region to be under represented in the LQLC regions because the coding region is generally more complex, which should make assembly more accurate. If the unreliable regions are enriched for calls in these

datasets, it may suggest that analysis of these regions produces a higher level of false-positives than the rest of the genome.

# MATERIALS AND METHODS

## Sample, Sequencing, and Alignment

Eight sets of paired-end, whole-genome, Illumina sequencing reads from a single sample from T. J. Tabasco, the sow whose genome was used to construct Sscrofa10.2, were used in this study[1]. BWA (v0.6.2: Li and Durbin, 2009) was used to align the reads to the Sscrofa10.2 reference assembly using default parameters. The reads were mapped to both the chromosomes and the unplaced scaffolds from the assembly. Any reads which mapped to chromosome Y were excluded as the sequences were from a female pig; consequently, we are unable to comment on the quality of the assembly of chromosome Y.

## Identifying Regions with Abnormal Coverage

SAMtools was used to filter the data to remove reads with a mapping quality less than 2 or which were improperly paired. BEDtools (v2.16.2: Quinlan and Hall, 2010) bamtobed was used to extract the chromosome, start positions and the end positions of whole sequencing fragments. BEDtools GenomeCov was then used to find per-base fragment coverage across the genome. BEDtools MakeWindows was used to make windows of 1000bp across the whole genome. Gap data was downloaded from the UCSC table browser (Karolchik et al., 2004) and BEDtools intersect was used to remove windows intersecting gap data. The median coverage for each 1000 base window across the genome was calculated. GC content is known to have a significant effect on coverage in sequencing methods that involve a PCR stage (Kozarewa et al., 2009). Coverage was normalized by GC content as described by Yoon et al. (2009). Briefly, the read coverage in each 1 Kb window (w) was adjusted by a multiplying factor *f*, with *f* equal to the ratio of the overall median across all windows divided by the median of all windows with the same GC percentage as that of the window *w*. Using the median instead of the mean prevented these values from being inflated by extreme outliers, as described by Zhang and Backstrom (2014). Any window with a normalized coverage over 55 or under 27 (2 *SD* from the mean; 41) was defined as having an abnormal coverage.

The removal of multimappers prior to coverage analysis may cause the detection of LC regions in certain sequence contexts in the genome that are more likely to contain multimappers (e.g., repetitive regions). Multimapped reads were extracted from the original bam file and read counts for these were calculated using Bedtools Coverage and the same 1000 bp windows used in the above coverage analysis; additionally raw read counts for each window were calculated in the same way from the original bam file. The percentage of reads in each window which were multimapped was calculated. Windows with >50% multimapped reads are likely to have been identified as LC due to the removal of these reads before coverage analysis. The regions with >50% multimappers were merged and intersect with the LC regions was calculated using Bedtools.

## Identifying Regions with Abnormal Insert Sizes

The mean and standard deviation of the insert sizes was calculated using Picard InsertSizeMetrics[2] (v1.113). Insert sizes were considered abnormal if they were more than 2 *SD* from the mean (427 bp). The merged BAM file was filtered for abnormally large (above 588 bp) and small (below 266 bp) insert sizes. BEDtools coverage was used to find the read count of the abnormal reads and the original BAM file using 1000 base windows with 200 overlap created with BEDtools MakeWindows. These data were used to calculate the percentage of abnormal reads in each window. A high proportion of small insert sizes was defined as a window with over 9.47% small insert sizes (2 *SD* above the mean of 4.22%) and a high proportion of large insert sizes was defined as a window with over 1.86% large insert sizes (2 *SD* above the mean of 0.12%).

## Identifying Regions with a Low Proportion of Properly Paired Reads

The mapped reads were filtered using SAMtools for the SAM flag 0×2, removing reads which were flagged as improperly paired. The percentage of properly paired reads was calculated as described for insert sizes. Any window with fewer than 70.59% (2 *SD* below the mean of 92.5%) properly paired reads was considered abnormal.

## Variant Calling

Single nucleotide polymorphism and indels were called using SAMtools mpileup, BCFtools and vcfutils varFilter. The resultant vcf file was filtered for homozygous variants, indicative of errors in the reference genome or sequencing errors. In order to include the entire regions covered by reads overlapping each variant, the regions spanning from 100 bases before to 100 bases after each variant were considered low quality.

## Merging

BEDtools was used to merge the regions identified by the above parameters into LQ, LC, and LQLC regions. BEDtools intersect was used to find regions of each group which overlapped with the coding region (regions downloaded from UCSC table browser; Karolchik et al., 2004). Sanger's gEVAL website[3] was used to inspect BAC and fosmid end alignments in a number of the identified regions.

## Assessing Effect of Identified Regions on Public Data

Known variant data were downloaded from dbSNP (Sherry et al., 2001) and the number of variants overlapping the abnormal regions were calculated. To assess the potential effect of these

[1]http://www.ebi.ac.uk/ena/data/view/ERP010190

[2]http://sourceforge.net/p/picard/wiki/Main_Page/

[3]http://geval.sanger.ac.uk/index.html

regions on WGS resequencing studies in pigs, the regions identified as CNVRs in Paudel et al. (2013) were downloaded and the number of regions overlapping the abnormal regions from the current study were calculated. Gene predictions based on RNA-seq data were downloaded from Ensembl (Cunningham et al., 2015) and the number of bases overlapping the identified regions calculated.

# RESULTS

## Alignment

582,271,856 reads mapped to the reference and 94.66% of these were properly paired (551,173,366 reads).

## Abnormal Regions

The effect of GC content on median coverage was as expected, with both high and low GC content regions having poor median coverage (**Figure 1A**).

While the coverage following GC normalization did follow a normal distribution, several extreme outliers inflated the mean and standard deviation. R (R Development Core Team, 2009) was used to find the mean and standard deviation of the majority of the data by overlaying a normal distribution on the data (**Figure 1B**). Using this method, we determined the mean coverage to be 41X and the standard deviation to be 7.

Regions identified by the parameters measured are summarized in **Table 1**. In total, 2.6% of the genome had abnormally high coverage, and 26.6% of the genome abnormally LC. Regions with a high percentage of fragment pairs with abnormally low and high insert sizes cover 3.99% and 1.52% of the genome, respectively. Regions with a low percentage of properly paired reads cover 4.95% of the genome. One of the largest regions identified (77.8 Kb) has abnormal coverage, insert sizes and read orientation (**Figure 2A**), and this is not uncommon, further examples are shown in **Figures 2B,C**.

There were a total of 62,463 regions with >50% multimappers and of these 99.3% overlapped with the LC regions. 66% of the regions identified as LC overlapped with the multimapped regions. The remaining LC regions had an unremarkable distribution of GC contents (data not shown) and the majority (81%) had 0 multimappers. The median read count per window for the whole genome was 264 and the median read count per window for the LC regions excluding those with >50% multimappers was 161.

We identified a total of 583,093 homozygous variants. Following merging, there were 245,972 regions identified as abnormal due to proximity to these variants covering 63,085,828 bases (2.25% of the genome).

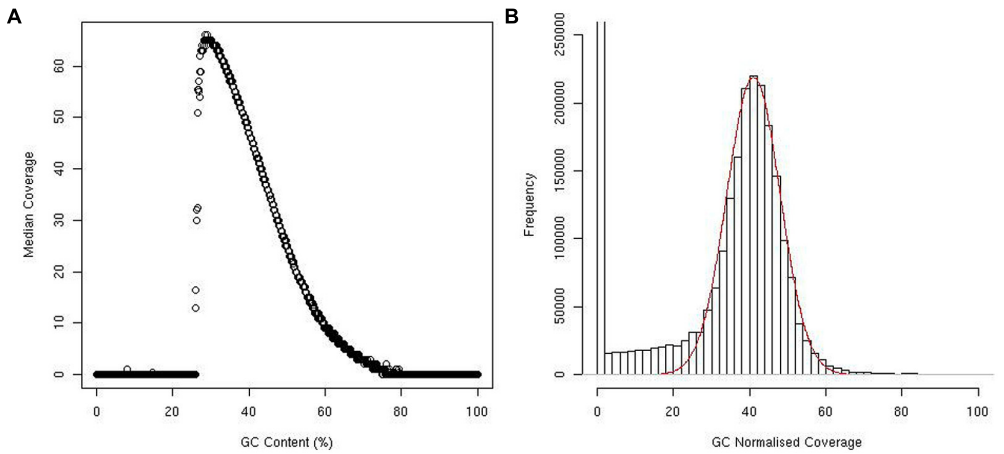



**FIGURE 1 | Plot showing median coverage of windows against percentage of GC content (A).** Histogram showing the distribution of window coverage, red line represents a normal distribution **(B)**.

**TABLE 1 | Table summarizing the regions identified by different parameters measured.**

| | No. of features | Mean feature size | Percentage of genome |
|---|---|---|---|
| High coverage | 60,281 | 1,202 | 2.6 |
| Small insert | 82,097 | 1,363 | 3.99 |
| Large insert | 31,833 | 1,343 | 1.52 |
| Improperly paired | 77,785 | 1,786 | 4.95 |
| Homozygous variants | 245,972 | 256 | 2.25 |
| Low quality (LQ) | 409,905 | 949 | 13.85 |
| Low coverage (LC) | 119,251 | 6,275 | 26.6 |
| **Total (LQLC)** | 337,276 | 2,753 | 33.07 |

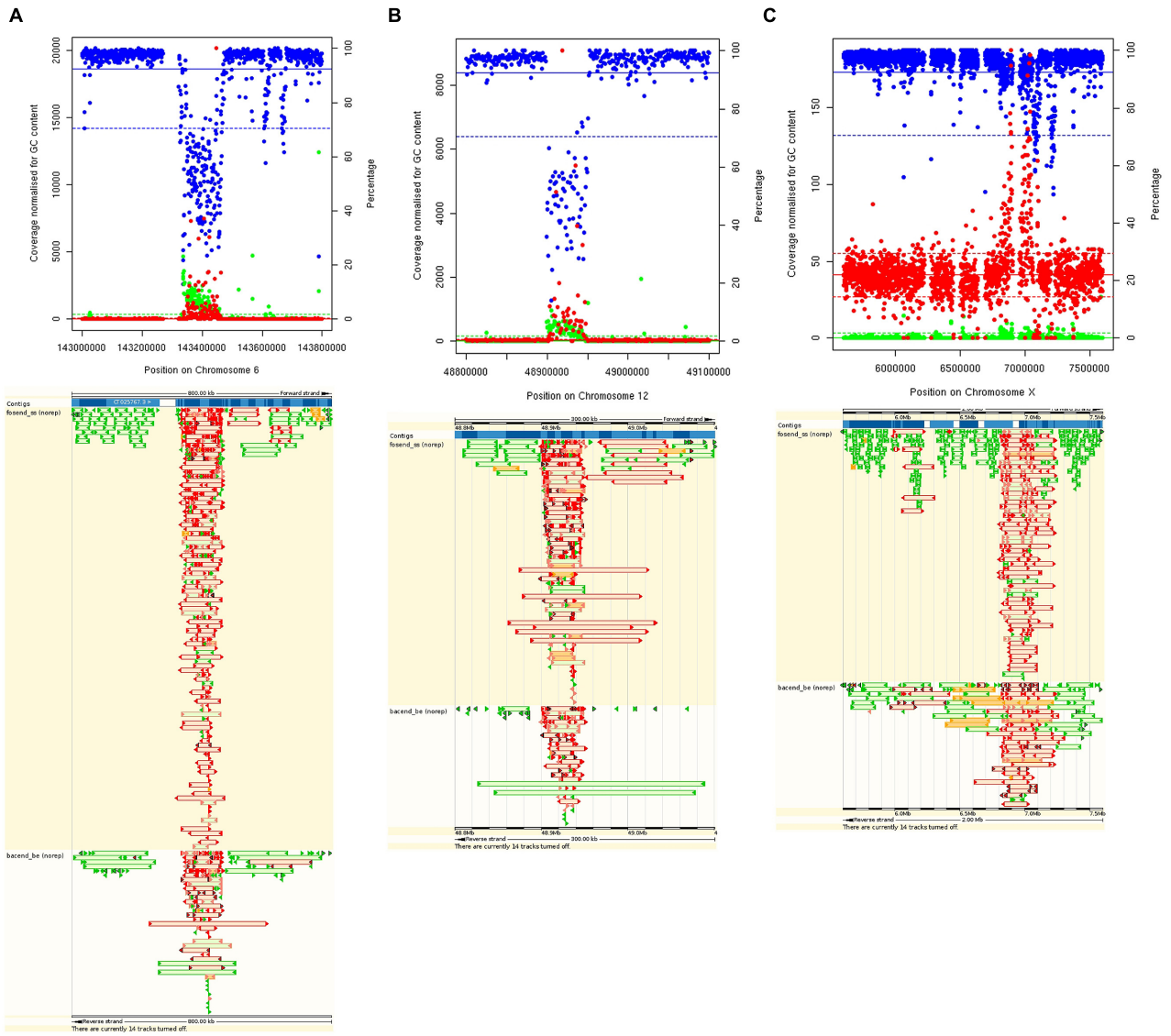



**FIGURE 2 | Plots showing examples of abnormal regions for multiple parameters on Chromosomes 6 (A), 12 (B), and X (C) (top).** GC normalized coverage (red) uses the left *Y* axis. Percentage of properly paired reads (blue) and percentage of high insert sizes (green) use the right *Y* axis. Means are represented by solid lines and 2 *SD* from the mean are represented by dashed lines. Bottom shows same regions viewed on the gEVAL browser with poorly mapped fosmids (top) and bac ends (bottom) shown in red.

## Merged Regions

After merging the regions with abnormal insert sizes, abnormal read orientation, and homozygous variant calls, we were left with 409,905 regions identified as being LQ, covering 13.85% of the genome.

In total, 337,276 regions were identified as being LQLC and the regions covered a total of 928,664,896 bases (33.07% of the genome). If the multimapped regions are excluded from the LC regions and the remaining LC regions are merged with the LQ regions these cover 17.3% of the genome.

The coding region data downloaded from UCSC table browser covered 587,219,382 bases (excluding chromosome Y) and of these 81,566,904 (13.89%) intersected with the LQ regions.

Of the coding region, 154,875,678 bases (26.37%) intersected with the LQLC regions.

## Impact on Public Data

The proportion of variants from publicly available data sets from Paudel et al. (2013) and dbSNP (Sherry et al., 2001) that fall in the abnormal regions are summarized in **Table 2**.

Paudel et al. (2013) identified 61,761 multi-copy regions (MCR), and from these identified 3,118 CNVRs. Of the CNVRs 1,081 (34.66%) lie in the LQ regions and 2,692 (86.3%) lie in the LQLC regions identified here.

The data downloaded from dbSNP (Release 104. Accessed: 05/05/2015) contain 52,634,111 known variants. In total,

**TABLE 2 | Table summarizing the proportion of called variants in publicly available data that fall in the abnormal regions identified in the current study.**

| | Total | LQ | LC | Combined (LQLC) |
|---|---|---|---|---|
| % of genome | – | 13.85% | 26.6% | 33.07% |
| % of coding region | – | 13.89% | 17.72% | 26.37% |
| dbSNP variants[a] | 52,634,111 | 19,121,760 (36.33%) | 15,483,445 (29.42%) | 27,009,232 (51.3%) |
| CNVRs[b] | 3,118 | 1,081 (34.66%) | 1,706 (54.71%) | 2,692 (86.3%) |
| RNA-seq genes[c] (intersecting bases) | 41,788,900 | 11,155,280 (26.69%) | 11,360,980 (27.19%) | 17,959,798 (42.98%) |

[a]*Data from dbSNP database (Sherry et al., 2001).*
[b]*Data from Paudel et al. (2013).*
[c]*Data from Ensembl (Cunningham et al., 2015).*

19,121,760 (36.33%) dbSNP variants were located in the LQ regions, 15,483,445 (29.42%) dbSNP variants were located in the LC regions and 27,009,232 (51.3%) dbSNP variants were located in the LQLC regions.

The gene predictions based on RNA-sequencing data covered 41,788,900 bases, 26.69% of these bases were in the LQ region (11,155,280), 27.19% were in the LC region (11,360,980) and 42.98% were in the LQLC regions (17,959,798).

## DISCUSSION

This work emphasizes the importance of accuracy in reference genomes in variant discovery research. Previous work by Zhang and Backstrom (2014) used sequencing reads from multiple chickens to detect misassemblies in the chicken genome. Here we used data from the same individual used to construct the pig reference assembly. We are therefore able to assess the assembly without introducing potential true variation that may be present by chance in multiple individuals; however, regions of the genome may have been incorrectly identified as low-quality due to true structural variation at heterozygous sites.

Regions of Sscrofa10.2 identified in this study were enriched for variants from dbSNP. The fact that the regions identified were enriched for variants in dbSNP, with the LQLC regions containing over half of the dbSNP variants, supports the assertion that these regions are enriched for false-positives; dbSNP contains large numbers of SNPs that are not validated and are potentially false-positives (Mitchell et al., 2004; Musumeci et al., 2010).

In the CNVR study by Paudel et al. (2013), 61,761 MCRs were identified and the authors state that the majority of these were common in all individuals sequenced; in this study 60,281 regions were identified as having high coverage and it is likely that there is overlap between these results. Studies looking for copy number gains may benefit from excluding the LQ regions from analysis. From the MCRs, 3,118 CNVRs were identified. The authors estimated that of these 2,664 (85.43%) were likely to be neutral or nearly neutral as they were common between different groups or were in non-genic regions, which is very similar to the number of CNVRs in the data that overlap the LQLC regions in the current study (2,692; 86.3%). CNVRs are called from sequencing data by comparison of read counts for a region with the average across the genome; it is likely that there are many false tandem repeats or collapsed repetitive regions in the assembly that would cause false copy number loss or gain calls. While regions identified as CNVRs are potentially variable regions between populations, breeds and individuals, calls based solely on comparison with the reference will give false-positives and false estimates of the copy numbers in true variable regions. Paudel et al. (2013) used copy number comparisons between individuals from different populations to identify MCRs that were variable between groups, which likely removed the majority of the false-positives. Other studies have used array-based methods to detect CNVRs in the pig genome (Chen et al., 2012; Wang et al., 2012) and of the regions identified in these studies, almost all of them fall in the LQLC regions (data not shown). This suggests these regions truly are enriched for CNVRs; however, enrichment of the unreliable regions for CNVRs may also suggest unreliable assembly around large duplications. In studies using whole genome resequencing, often small sample sizes are used and too much confidence may be given to the reference. It would be advisable in studies using Sscrofa10.2, and references of other species that may contain similar inaccuracies, not to call CNVRs based solely on comparison with the reference, but from regional variation in read count between individuals as has been done previously for genomes which lack a reference following co-assembly (Nijkamp et al., 2012) and when comparing sequences from cancer cells to healthy cells (Chiang et al., 2009; Koboldt et al., 2012). Similarly, researchers using other techniques that rely on counting reads mapped to the reference genome such as ChiP-seq and RNA-seq should be aware that these errors may cause inaccurate calling or expression estimates. In RNA-seq, read counts are used to estimate expression levels; unexpected CNV between the reference and the sample sequence could cause over- or under-exaggerated read counts, potentially resulting in false-positives or false-negatives. RNA-seq is prone to off-target mapping (Mortazavi et al., 2008), particularly at higher depth (Tarazona et al., 2011); true peaks can often be distinguished from off-target mapping using an expression threshold. However, misrepresentation of the copy number of a region in the reference assembly may exaggerate off-target peaks above the threshold and cause false-positives, exaggerate true peaks causing inaccurate expression estimates, or reduce true peaks causing false-negatives or underestimation of expression. The regions identified here were enriched for RNA-seq gene predictions, more so than the annotated coding region, which may suggest an increased false-positive rate in these regions from this method.

A large amount of the genome showed LC. While these regions may suggest errors in the reference genome, such as false tandem duplications (Zhang and Backstrom, 2014), they do so with less confidence than the other parameters measured. The study by Paudel et al. (2013) reported a considerable number of copy number losses and subsequently excluded these from further analysis as they were likely enriched for false-positives; the fact that this excess of LC regions has been encountered by other researchers may suggest that the problem is with the quality of the genome assembly or region mappability rather than the quality of the data used in the current study. Regions with LC were analyzed separately as LC may be an indicator of the quality of the sequencing data, PCR bias or poor mappability and not necessarily inaccuracy in the reference. The majority of the LC regions were explained by their large proportion of multimappers; the regions were identified as LC because multimappers were excluded from the coverage analysis. These regions may not be misassembled, but rather of poor mappability due to, for example, low complexity or repetitive sequences. Of the LC regions which were not explained by multimappers there was no evidence of extreme GC content causing the reduced coverage and the majority contained no multimappers; the LC in these regions likely relates to misassembled areas in the reference genome, or potentially heterozygous structural variants in the individual. Where the LC is explained by poor mappability, it may still be advisable to exclude these regions from SNP and indel analyses as this is likely to yield LQ variants with a high rate of false-positives. Studies requiring identification of only the highest quality variants would reduce computational burden and false-positive rate by excluding the LC regions. In studies more concerned with finding variants relating to a specific phenotype, if LC regions are included, variants identified in them may be treated as low-confidence, but not necessarily excluded entirely. The percentage of dbSNP variants in the LC region is not as high as in the LQ region, however, fewer variants may be called in poor mappability regions due to the common practice of filtering out low mapping quality reads before proceeding to variant calling, reducing depth and subsequently the chances of calling a variant in these regions. The proportion of the genome identified here as LQ is likely to be an over-estimation of the proportion that is misassembled. The individual may have true, heterozygous structural variation that cause some of these regions to appear misassembled and this analysis has been intentionally strict to allow downstream bioinformatic analysis to focus on only the highest confidence regions of the genome by excluding LQLC regions. The number of variants identified in studies employing variant calling is often extreme and strict filtering techniques are employed to reduce the number to a more tractable level for validation (MacArthur et al., 2012; Ai et al., 2015). Excluding regions which are likely to be enriched for false-positives may significantly reduce computational burden and increase accuracy. Strict filtering after variant calling may cause the loss of variants of interest and it is desirable to reduce the initial number of variant calls as much as possible to reduce the need for excessive filtering. While variants of interest may lie in the low-confidence regions identified here, the excess of false-positives in the region make it unlikely that they will be easily identified. However, discovery of variants outside of these regions will benefit from the reduced number of false-positives in the dataset. Many variant callers and filtration methods will consider depth and mapping quality and are likely to exclude a number of false-positive variants from these regions by default; however, computational burden would be decreased by excluding unreliable regions, which will be particularly relevant with large datasets. Other methods that use regional read count data need to be aware that Sscrofa10.2 does not accurately represent the copy number of alleles in the original Duroc sow's genome. Clearly in studies searching for CNVRs, excluding the LQLC regions, which are potentially enriched for true CNVRs, is not an option. In such studies it would be beneficial to compare individuals in a study with one another rather than with the reference, as is done in somatic variant calling comparisons between healthy cells and cancer cells (Roberts et al., 2013), to filter out variation that is common in all individuals, or to exclude the LQ regions only. The degree to which misassemblies will affect research results depends on a number of factors including the tools used, the type of misassembly and the type of analysis; for example, the incorrect order of contigs will negatively affect read-pair mapping and collapsed duplications may cause incorrect calling of SNPs – though SNP callers may accurately filter many of these. Similar inaccuracies to those found here are likely to be present in the reference genomes of other non-human species. With the price of sequencing continuing to fall, the number of large-scale sequencing studies on species with draft genomes will undoubtedly increase; awareness of inaccuracies in these references will decrease computational burden and increase accuracy. Identifying regions that are inaccurate and producing new, more accurate assemblies will greatly increase the power of whole-genome resequencing studies in non-human species.

## Availability of Data

The regions identified in this study have been made available as three bed files: LQ regions, LC regions, and LQLC regions. BED files are available to download from http://www.ark-genomics.org/outputs/identification-low-confidence-regions-pig-reference-genome-sscrofa102

## FUNDING

This work was funded by grants from the Roslin Foundation, Genus Plc and by a BioSciences KTN CASE studentship. The work was enabled by funding from the Biotechnology and Biological Sciences Research Council including Institute Strategic Programme and National Capability grants (BBSRC; BBS/E/D/20310000, BB/J004243/1). DNA sequencing was provided by Edinburgh Genomics (http://genomics.ed.ac.uk). Edinburgh Genomics is partly supported through core grants from the National Environmental Research Council (NERC R8/H10/56), Medical Research Council (MRC MR/K001744/1) and The Biotechnology and Biological Sciences Research Council (BBSRC BB/J004243/1).

## REFERENCES


Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet* 47, 217–225. doi: 10.1038/ng.3199

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18–R18. doi: 10.1186/gb-2011-12-2-r18

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., et al. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5473–5478. doi: 10.1073/pnas.1418631112

Bianco, E., Nevado, B., Ramos-Onsins, S. E., and Pérez-Enciso, M. (2015). A deep catalog of autosomal single nucleotide variation in the pig. *PLoS ONE* 10:e0118867. doi: 10.1371/journal.pone.0118867

Chen, C., Qiao, R., Wei, R., Guo, Y., Ai, H., Ma, J., et al. (2012). A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics* 13:733. doi: 10.1186/1471-2164-13-733

Chiang, D. Y., Getz, G., Jaffe, D. B., O'kelly, M. J. T., Zhao, X., Carter, S. L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103. doi: 10.1038/nmeth.1276

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. *Nucleic Acids Res.* 43, D662–D669. doi: 10.1093/nar/gku1010

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi: 10.1038/nature11622

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet. Adv.* 47, 435–444. doi: 10.1038/ng.3247

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496. doi: 10.1093/nar/gkh103

Kelley, D. R., and Salzberg, S. L. (2010). Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol.* 11, R28–R28. doi: 10.1186/gb-2010-11-3-r28

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nat. Methods* 6, 291–295. doi: 10.1038/nmeth.1311

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 335, 823–828. doi: 10.1126/science.1215040

Mitchell, A. A., Zwick, M. E., Chakravarti, A., and Cutler, D. J. (2004). Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 20, 1022–1032. doi: 10.1093/bioinformatics/bth034

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Musumeci, L., Arthur, J. W., Cheung, F. S. G., Hoque, A., Lippman, S., and Reichardt, J. K. V. (2010). Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.* 31, 67–73. doi: 10.1002/humu.21137

Nijkamp, J. F., Van Den Broek, M. A., Geertman, J.-M. A., Reinders, M. J. T., Daran, J.-M. G., and De Ridder, D. (2012). De novo detection of copy number variation by co-assembly. *Bioinformatics* 28, 3195–3202. doi: 10.1093/bioinformatics/bts601

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L. A. F., Bosse, M., Bastiaansen, J. W. M., et al. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14:449. doi: 10.1186/1471-2164-14-449

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing [Online]*. Vienna: The R Foundation for Statistical Computing. Available: http://www.R-project.org/ [Accessed May 01, 2015].

Robert, C., Fuentes-Utrilla, P., Troup, K., Loecherbach, J., Turner, F., Talbot, R., et al. (2014). Design and development of exome capture sequencing for the domestic pig (*Sus scrofa*). *BMC Genomics* 15:9. doi: 10.1186/1471-2164-15-550

Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., et al. (2013). A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* 29, 2223–2230. doi: 10.1093/bioinformatics/btt375

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21, 2213–2223. doi: 10.1101/gr.124321.111

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., et al. (2005). Fine-scale structural variation of the human genome. *Nat. Genet.* 37, 727–732. doi: 10.1038/ng1562

Wang, J., Jiang, J., Fu, W., Jiang, L., Ding, X., Liu, J.-F., et al. (2012). A genome-wide detection of copy number variations using SNP genotyping arrays in swine. *BMC Genomics* 13:273. doi: 10.1186/1471-2164-13-273

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Zhang, Q., and Backstrom, N. (2014). Assembly errors cause false tandem duplicate regions in the chicken (*Gallus gallus*) genome sequence. *Chromosoma* 123, 165–168. doi: 10.1007/s00412-013-0443-8


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer Martien Groenen declares that, despite collaborations with the author Alan L. Archibald, the review process was handled objectively.

## 2.3 Conclusion

The analysis to identify potentially misassembled or unreliable regions in the draft pig reference genome (Sscrofa10.2) identified a very large percentage of problem regions. These regions can now be used as a filter to reduce false-positives in datasets from NGS in pigs when the Sscrofa10.2 assembly is used as the reference. This work suggests that there are major flaws in this widely used reference genome. The overall impact of these flaws is not clear, but they may have obscured phenotypically relevant regions in past research.

The use of SNP chip data in animal breeding, for example in genomic selection, is generally robust against small errors in the genome. However, as exome and whole genome sequence (WGS) data are increasingly used in genetics studies or applications, the importance of the reference genome's accuracy increases. Applications in animal breeding include the use of WGS (Daetwyler et al., 2014), genotyping-by-sequencing (De Donato et al., 2013), and imputation of genome sequence information from SNP genotypes (Marchini et al., 2007). Additionally, with increased use of technology such as CRISPR gene editing in pigs (Whitworth et al., 2015, Wells and Prather, 2017, Burkard et al., 2017, Burkard et al., 2018), the reference must be reliable both in terms of completeness and accuracy in order to design guide RNA and reduce the chance of off-target effects.

This work highlights the fact that reference genomes are not necessarily reliable. Many workflows assume that the reference genome is a truth against which to compare other individuals, however it is usually neither an

accurate representation of a polyploid species, nor of the haploid genome. In the case of the pig, the reference genome was assembled using a method considered the gold standard at the time, yet it has an abundance of structural errors. What level of accuracy then can be expected of assemblies from WGS shotgun Sanger or Illumina data?

The methods described here are also applicable to the improved reference genome sequences being developed for pigs and other farmed animals. These new reference genomes have greatly improved contiguity they have been assembled primarily from third generation long read sequencing technology. As the raw sequence data from these technologies has a high error rate, the filtering methods described here remain relevant.

The work in this chapter has been cited 22 times (as of 18/07/2018; Google Scholar). This includes the use of the defined LQLC regions to filter datasets for low-quality structural variants in both European (Keel et al., 2017) and Chinese (Yang et al., 2017) pig breeds. The methods described for adjusting short-read coverage for GC content have been used to estimate the size of the genomes of two butterfly species, *Heliconius melpomene* (Davey et al., 2016) and *Leptidea sinapsis* (Talla et al., 2017). In both cases the genome size estimate agreed with estimates from other methods, and in the case of *H. melpomene*, revealed that most of the sequence missing from the assembly was in collapsed repeats of sequence already represented in the assembly that could likely be resolved from the short-read data alone. Additionally, Wu et al. (2017) included use of the GC content adjustment and coverage limits defined in this chapter to identify errors in genotyping from

short-read data. The coverage limits identify likely false heterozygote calls, where unexpectedly high coverage over a heterozygous site may indicate a CNV or paralogous site.

The pig is an important agricultural species, and it is likely that the falling cost of sequencing will lead to an increase in sequencing data produced by the community for this species. For these data to be used to their full potential the limitations of the reference genome must be taken into account.

# CHAPTER 3: IDENTIFICATION OF PROTEIN-TRUNCATING VARIANTS AND REGIONS ASSOCIATED WITH REPRODUCTIVE SUCCESS IN PIGS

---

*"I don't know what I'm looking for."*
*"Why not?"*
*"Because… because… I think it might be because if I knew I wouldn't be able to look for them."*
*"What are you, crazy?"*
*"It's a possibility I haven't ruled out yet."*
-Douglas Adams, The Hitchhiker's Guide to the Galaxy

## 3.1 Introduction

The relatively short gestation time of ~114 days and the large litter sizes of pigs allow for efficient promotion of beneficial traits through artificial selection (Zak et al., 2017). Selection can be done based on the performance of the individual, or more commonly the performance of close relatives and progeny of the individual. Variability of traits within a herd allow for selection of the most successful individuals for breeding, leading to improvement of the performance of the population as a whole. Commercial pig breeding populations usually have a pyramidal population structure (figure 3-1), with a nucleus herd of a relatively small population of high performers which feeds into multiplier farms, and eventually commercial farms for pork production (Visscher et al., 2000, Shepherd and Kinghorn, 1992). Genetic variants that cause reduction in reproductive performance circulating in nucleus herds can have an impact on the progress of improvement in these herds. Improvements to important agricultural traits in the nucleus herd can take years to reach the commercial farms at the bottom of the pyramid.

Selection in the nucleus herd is largely driven by information from an individual's genotype and phenotypic history, including the success of their progeny. Discovery of phenotype associations and the process of SNP-chip based selection itself depends on the presence of haplotypes that contain both a marker gene and a causative locus. The number of haplotypes in a nucleus herd is difficult to determine as it is a function of the diversity of the population and the typical length of the haplotypes

and will be specific to the herd. There are typically several thousand individuals in a nucleus herd, however the number of haplotypes also depends on the size of the founding population and how long the population has been established, with the number of haplotypes typically being fewer in young populations and populations with small founding populations and therefore high inbreeding. This is further complicated by migration in and out of the population which will have an impact on the number of haplotypes. Importantly, however, the majority of haplotypes will be common, with new haplotypes being formed relatively infrequently. This has been proposed as a benefit for the whole genome sequencing entire populations as the sequences of the majority of individuals can be imputed by sequencing a few individuals specifically targeting focal individuals that carry the rarer haplotypes (Gonen et al., 2017, Hickey, 2013).

Reproductive success can be influenced by a number of genetic factors, including factors affecting the male reproductive system, the female reproductive system or the development of embryos. In the male, chromosomal defects or smaller defects in specific genes may impact on numbers of sperm cells in the ejaculate, ejaculate volume and sperm motility (Zak et al., 2017, Smital et al., 2005). In females, ovulation rate, age at puberty, wean-to-oestrus interval, number stillborn, litter size and number of

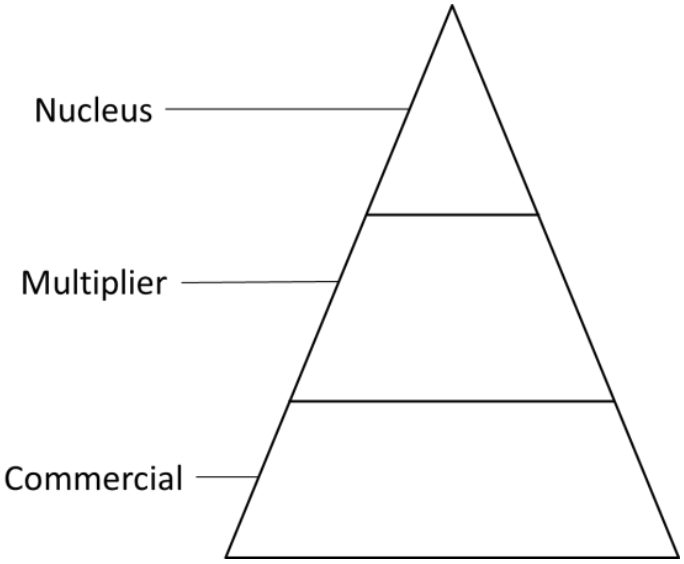



**Figure 3-1- Pyramidal structure of animal breeding programs**

piglets surviving to weaning are common phenotypes targeted by selection (Zak et al., 2017), but these traits are generally controlled by many genes and molecular mechanisms behind these traits are complex. Litter size may be a phenotype of the female associated with ovulation rate, immunological factors, or uterine quality and capacity. However litter size can also be reduced by genetic defects of foetuses inherited from one or both parents that impact on survival. This includes variants that are incompatible with life, which may cause death before implantation, after implantation, or post-parturition. Additionally, variants in the dam and/or the offspring that might reduce the success of implantation and the efficiency of the maternal/foetal interface may cause death or stunted growth of foetuses *in utero* as has been observed in humans (Pardi et al., 2002, Burton et al., 2009, Chelbi et al., 2012, Shan et al., 2015, Parle-McDermott et al., 2005). Survival to weaning is an even more complex trait, including factors such as litter size, farrowing survival and mothering ability of the sow (Zak et al., 2017).

Ideally, when specific variants that directly cause a detrimental trait can be identified and selected against, they may be purged from the population. More commonly, marker variants physically close to the causative variant, but not causative themselves can be identified and used for selection as a genotype “linked” to the causative variant, the effectiveness of which depends on the level of linkage disequilibrium (LD) between the marker and the causal genetic variant (van den Berg

et al., 2016) and the complexity of the phenotype. Breeding based on genetic markers alone can be successful in reducing prevalence of an unwanted trait, but this may not completely purge any causative genetic variants due to recombination events creating new haplotypes and disassociating the variants (Morrell et al., 2006). While causative variants may be identified by investigating candidate genes from knowledge in other species or candidate regions from association analyses, sequencing of the genome or exome can be used to identify likely candidates without such prior knowledge. Additionally, while causative variants circulating in one line of commercial pigs are unlikely to be found in another line, identifying the genes involved provides candidate genes for other lines which may have a different variant affecting the same gene or molecular pathway.

Whilst most of the traits of interest to the pig industry are controlled by many genes each with small effects on the trait, genetic variants in a single gene can have major deleterious effects, including compromised development and lethality. Examples of such major gene effects are the monogenic inherited diseases in humans. For many of the common human inherited diseases, the genes involved have already been identified. Often the causal genetic variant for such monogenic inherited diseases can be classified as Loss of Function (LoF) variants and include frameshifts or premature stop codons that result in alleles that no longer encode a functional protein. Putative LoF variants can be identified in genome or exome sequence data. Pigs are highly similar to

humans in a number of ways, as discussed in the literature review of this thesis, and are of interest for use as medical models. Thus, scanning pig exome and genome sequence data for putative LoF variants, especially in genes already associated with a human inherited disease, may not only address pig industry issues such as embryo survival, but also potentially yield pigs with relevant genotypes as animal models for human diseases. These individuals may be of use in research to understand the mechanisms, manage, and potentially cure these diseases in humans (Prather et al., 2013, Perleberg et al., 2018).

The exome sequencing of 96 boars by Robert et al. (2014) identified over 250,000 variants after hard filtering, of which over 38,000 were predicted by Ensembl VEP (McLaren et al., 2016) to alter a protein. While variant annotation can help to identify variants that might cause an observable phenotype, on its own it still produces a list of thousands of candidates that may be of interest and it cannot further prioritise the variants. The majority of the variants identified are likely to be false-positives or benign variants, the task of reducing the variants to a manageable list of prioritised candidates remains with the researcher.

In human research, in cases where the cause of a highly specific phenotype is sought, often candidate genes are investigated first. These are genes that have been associated with similar phenotypes in other individuals or species, or those located within a region flagged by another method such as a genome wide association study (GWAS). In the case of the pig exomes, relatively non-specific phenotypes are

under investigation with only the aim of improving reproductive efficiency in pig breeding. These complex phenotypes may be strongly influenced by the environment and likely involve interplay between multiple genes. In order to find candidates for these phenotypes, variants must be filtered based on their likely impact on the coded protein and the importance of that protein to the phenotype. For example, the loss of a single amino acid from the end of the protein is unlikely to render it non-functional, and a disruption in genes from a multigene family with redundant functions or with non-essential functions, such as olfactory receptors, are unlikely to cause lethality.

As previously discussed in Chapter 1 of this thesis, within the protein-coding region, protein truncating variants are considered the most likely to cause an observable phenotype. As protein truncating variants disrupt the protein's reading frame, their position within the protein coding region of a gene will determine how much of the protein is lost, so protein truncating variants that occur early in the coding sequence are likely to have the largest effect on the protein's function. Additionally, variants that are observed in the homozygous state in healthy adults cannot be the sole cause of embryonic lethality.

While bioinformatic approaches to identify underrepresented variants that may truncate a protein can produce useful candidates, without further phenotypic information it cannot reveal whether they are associated with an observable detrimental or beneficial effect. A common method for locating variants associated with a trait in livestock

is a GWAS. GWAS involve looking for evidence of associations between genetic variants and variation in the trait of interest. Such GWAS typically involve the use of SNP arrays to genotype a large number of SNPs fairly evenly dispersed across the genome for individuals with phenotypic data which can be associated with the genotypes. GWAS does not directly identify causative variants, but relies on LD between loci in the genome causing a number of variants to coincide on a haplotype with the causative variant. Identifying significantly associated SNPs narrows the search space from the whole genome to one or more loci, the size of which depends on the number of animals (or more accurately the number of informative meioses) in the study, the number of informative genetic markers (typically SNPs) and the level of LD in the population. The success of GWAS depends on how many loci are involved with the phenotype, the effect size of each locus, allele frequencies, the sample size, the SNP array used and how heterogeneous the phenotype under investigation is in the population (Visscher et al., 2017). GWAS typically involve thousands of individuals to increase the power of the analysis (Spencer et al., 2009). Often, to increase the number of individuals in the analysis while keeping costs low, a large population is genotyped with a low density SNP array and a subset of the same population, the reference population, is genotyped on a higher density SNP array. Using the data from the reference population, haplotypes in the population can be identified and genotypes for the SNPs missing from the lower density

array can be imputed (Wu et al., 2016, Habier et al., 2009, Weigel et al., 2009, Wiggans et al., 2012, Corbin et al., 2014, Bolormaa et al., 2015). While imputation reduces the accuracy of the GWAS, this is usually acceptable to achieve the increase in power from the larger pool of individuals. Imputation accuracy depends upon the relatedness of the individuals, the extent of linkage disequilibrium in the population and the correct ordering of the genetic markers in the genome, i.e. the quality of the reference genome. In contrast, the correct ordering of the markers is not essential for the association analysis, although errors make interpretation of the results more challenging.

A benefit of GWAS is that it can take all variants into account including those predicted to be of lower consequence by variant annotators, whereas the filtering of exome variants alone must first focus on only those with the highest predicted consequence to produce a manageable list of candidates to follow up on. GWAS enables potentially causative missense variants and non-coding variants that might affect splicing or the regulation of expression to be investigated. Additionally, if variants that have been predicted to have a high impact on the protein also appear to have an association with a phenotype, this increases the confidence that the candidate may be causative of an observable phenotype. However, both imputation and GWAS are unreliable where variants have a low minor allele frequency (MAF), with lower frequency variants being more prone to inaccurate imputation (Zheng et al., 2015) and misleading GWAS results (Tabangin et al.,

2009). In a commercial herd under strong selection, we would expect a low MAF for variants with a strong detrimental phenotypic effect due to selective pressures against undesirable phenotypes, but variants with a more quantitative effect may be more common. By both identifying rare protein truncating variants and performing GWAS to identify variants associated with traits, both common and rare candidates can be prioritised for further investigation.

In this chapter, filters are placed on the variants called from 96 pig exomes to create a short list of candidate variants on which to focus further work with emphasis on those that may cause reduced reproductive success or embryonic lethality. Additionally, variants in genes known to be associated with human disease are identified for their potential as use for naturally occurring medical models. In order to identify associations between variants and two reproductive phenotypes, imputation from 60K SNP chip to the exome variant data set is used followed by GWAS. The majority of the work in this chapter was done by the author, apart from additional sequencing of candidate variants from 27 dams following DNA extraction & whole genome amplification which was carried out by Jennifer Dzelil at Genus Pig Improvement Company (PIC).

## 3.2 Methods

### 3.2.1 Datasets

Exome sequencing data from 96 purebred boars from a commercial synthetic line were the primary dataset in this chapter. The data are those presented by Robert et al. (2014) in their paper describing the development of a porcine exome capture platform. These data had been produced from exome sequence data mapped to Sscrofa10.2 using BWA aln (Li and Durbin, 2009) and variant calls were made using GATK UnifiedGenotyper (McKenna et al., 2010) recommended practices including the use of the following tools: RealignerTargetCreator, IndelRealigner, FixMateInformation, BaseRecalibrator and UnifiedGenotyper. Base recalibration was done using Ensembl release 72 as known sites. Hard filters already applied to the data using GATK VariantFiltration as described in the paper are: remove variant calls located within a cluster where three or more calls are made in a 10 bp window; remove if there are at least four alignments with a mapping quality of zero (MQ0) and if the proportion of alignments mapping ambiguously corresponds to 1/10th of all alignments; remove variants which are covered by fewer than 5 reads; remove variants with a quality score below 50; remove variants with low variant confidence over unfiltered depth of non-reference samples (QD<1.5); filter based on strand bias using Fisher's exact test: FS > 60.0 for SNP calling, FS > 200.0 for indel calling.

In addition to the exome sequencing data, PorcineSNP60v2 BeadChip (Illumina) genotyping data for 351 individuals from the same line were provided by Genus PIC as well as two estimated breeding values (EBVs) for all 447 pigs for two phenotypes: number born in a purebred cross (Born to purebred; BTP) and number stillborn (NSB). Genus PIC use an internally developed two-trait multivariate mixed model augmented to predict random animal effects for animals without recorded phenotype data. The more data available to predict an animal's EBV, the more reliable the EBV is, with reliability scores ranging from 0 to 1. Both of these EBVs use the fixed effects of farm, week of farrowing, year of farrowing, parity, random animal effect, and a random repeat effect of the sow. Details of pedigrees were also provided for the 447 pigs detailing any $1^{st}$ and $2^{nd}$ generation relationships between individuals. While a larger dataset than this one with more individuals and a wider range of phenotypes could provide more accurate and detailed results, the pigs are from a commercial line and the data access was restricted accordingly.

As the initial set of 96 exomes were not chosen for extreme phenotypes and comprise a relatively small group of individuals, these will be used as a reference population to impute exome variants for the additional animals for which EBVs and SNP chip genotypes are available so that a GWAS can be performed for the two EBVs.

### 3.2.2 Bioinformatic filtering and validation

Unless stated otherwise, tools were run using default parameters. Ensembl VEP (v.80; McLaren et al., 2016) was used to annotate the exome variants.

The first filter applied to the exome variant data set was the removal of variants falling within the LQLC regions described in Chapter 2 of this thesis. These were filtered from the vcf using BEDtools intersect (v.2.16.2; Quinlan, 2014). GATK (v.2.3.9) VariantEval was used to calculate the transition/transversion (Ts/Tv) ratios for the original set of variants, those that were excluded and those that were not. The Ts/Tv ratio for coding exons in humans has been reported to be close to 3:1 (Heinrich et al., 2013, Mistry et al., 2015), so a change in Ts/Tv towards 3:1 may suggest an increase in accuracy.

The output from Ensembl VEP allows for the selection of variants that cause a stop gain or a frameshift (the major classes of protein truncating variants) or otherwise flagged as high impact and to select for those variants which occur in an exon that is within the first 50% of the total exon number for the gene or variants in genes where there is only one exon.

VCFtools (v. 0.1.12a; --hardy tool; Danecek et al., 2011) was used to calculate the expected Hardy-Weinberg Equilibrium (HWE) of the frameshift and stop gain variants and to find whether the observed genotypes were significantly different from this (exact test, $p<=0.05$)

Those that deviated from the expected HWE and had no alt homozygotes were prioritised and literature searched for relevant phenotypes. These variants were enriched for olfactory receptors, as these were unlikely to be relevant to the phenotype of interest and tend to be redundant in function, not strongly conserved and in poorly assembled gene clusters they were excluded at this stage. The HWE filter produced a very short list and was not used to exclude any variants from the main dataset.

An additional filter was applied to remove variants where fewer than 70% of the individuals carrying the variant had at least 40X coverage over the variant position. The remaining variants were prioritised by removing variants with no gene name in Ensembl or NCBI's annotation of Sscrofa10.2 and no clear ortholog in Ensembl. Variants with homozygotes for the alternative allele present were also removed.

To further validate prioritised variants, a 6-frame translator (https://web.expasy.org/translate/) was used to confirm the consequence predicted by Ensembl VEP. Read alignments for these variants were visualised using IGV (v2.3.88; Thorvaldsdóttir et al., 2013) to confirm that these agreed with the variant call.

A literature search was done to identify genes associated with common human diseases including cystic fibrosis, haemophilia, muscular dystrophy, sickle cell anaemia, Angelman syndrome, phenylketonuria, polycystic kidney disease, Tay-Sachs, spinal muscular atrophy type 1,

factor V Leiden thrombophilia and Graucher's disease, and the unfiltered exome variant set was searched for variants in these genes that were ranked deleterious according to SIFT (Ng and Henikoff, 2003). Two missense variants in CFTR, the gene associated with cystic fibrosis, were added to the short list of variants of interest due to their potential use in medical models if an altered phenotype is present.

Two further variants were selected for validation despite not meeting the above criteria. The variants in IL1RAP and TGFB2 do have homozygotes present in the population. As the individuals sequenced are adult pigs, these cannot be the sole cause of lethality, however based on previously published work they may have an impact on economically important phenotypes, namely embryo implantation and immune response (Mathew et al., 2016, Orelio et al., 2008, Dinarello, 2000), and growth and development (Memon et al., 2008).

### 3.2.3 Variant validation by sequencing

All of the lab work in this section was carried out by the author at the ABS/Genus lab in Wisconsin, USA.

#### *3.2.3.1 Primer design, resuspension and sample selection*

In order to validate the list of 21 prioritised variants, primers were designed for 20 regions (one pair of variants, the missense SNPs in CFTR, were within a short distance of one another and could be captured in a single amplicon) in order to amplify and resequence the

variants to confirm the initial variant discovery. Primers were designed using the BLAST primer design tool (Ye et al., 2012) aiming for a PCR product >500 bp long. The list of 96 PIC boars was reduced to a set of 20 boars that included at least one individual heterozygous for each of the variants. The primers were ordered from Integrated DNA Technologies (Iowa, USA) and upon delivery were resuspended in UltraPure Distilled Water (Invitrogen; California, USA) to make a stock dilution of 100 µM. Each tube was thoroughly vortexed and allowed to stand at room temperature for at least 10 minutes before re-vortexing to ensure resuspension. Working stock was diluted by transferring

10 µl of stock dilution to a new 1.7 ml microcentrifuge tube, adding 90 µl of UltraPure Distilled Water and vortexing thoroughly. Primers were stored at 4ºC prior to PCR.

### *3.2.3.2 DNA extraction*

For each of the 20 boars, a small piece of tissue, roughly 0.3-0.4 $cm^3$ was cut off either an ear or tail tissue sample that had been stored at -20ºC. The tissue was shredded with a clean razor and placed in a 1.7 ml microcentrifuge tube. Extraction was carried out on the Qiagen (Hilden, Germany) EZ1 Advanced using an EZ1 Advanced DNA Tissue Card. To prepare for this, 190 µl of Buffer G2 was added to each sample, followed by 10 µl proteinase K. These were then incubated in a water bath at 56ºC overnight to lyse the tissue. 200 µl of lysate was transferred into a 2 ml screw cap tube, discarding any remaining solid

tissue. Loading of the machine was done according to manufacturer's instructions and the DNA was eluted into 100 µl elution buffer. The DNA was stored at 4ºC prior to PCR.

### *3.2.3.3 PCR*

PCR was carried out for between 1 and 4 DNA samples per primer pair. Where a variant was carried by only one individual, this individual was amplified in duplicate, where a variant was carried by 2-4 individuals, all individuals were used, where a variant was carried by more than 4 individuals, a random set of four carriers were selected. A master mix was made using 5 µl of Long Amp Buffer (New England Biolabs; Massachusetts, USA), 1 µl dNTPs, 16.75 µl UltraPure Distilled Water, and 0.25 µl Long Amp (New England Biolabs) per well to be used plus 2 to allow for pipetting error. The mix was vortexed and briefly spun down in a centrifuge. PCR plates were arranged to have one pair of primers per column and to each well 23 µl of master mix and 0.5 µl each of the appropriate forward and reverse primer were added. 1 µg of DNA was added to the wells in the rows of the primer pairs according to which individuals carry the variant. The PCR plates were sealed with Adhesive BioRad (California, USA) Microseal 'B' PCR plate seals. The PCR plates were placed in GeneAmp PCR System 9700 thermocyclers (Applied Biosystems; California, USA) with the following conditions: 94ºC for 5 minutes, 35x(94ºC for 30 seconds, 58ºC for 30 seconds, 65ºC for 1 minute 30 seconds) 65ºC for 7 minutes, hold at 4ºC. PCR product was stored at 4ºC.

### *3.2.3.4 Gel electrophoresis*

3% agarose gel was prepared by mixing 18 g agarose (VWR Biotechnology; Pennsylvania, USA), 60 ml 5xTBE and 540 ml pure $H_20$, swirling the bottle to mix, and microwaving for around 10 minutes on 50% power, stopping to swirl again at 5 minutes. The mixture was allowed to cool slightly (~5 minutes), then 16 µl 1% Ethidium Bromide was added and the mixture was swirled again. The gel was poured into a large gel tray with three rows of 40 well combs and left to set for ~40 minutes. 10 µl of each PCR product and 2 µl Amaranth loading dye (Sigma; Missouri, USA) were mixed by pipetting before being added to the wells. 3 µl of 100bp ladder (New England Biolabs Quick-Load 100bp) and 1 Kb ladder (New England Biolabs Quick-Load 1Kb) were added to the two leftmost wells and two central wells of each row. The gels were run at 50 V for 3 hours on the BioRad (California, USA) PowerPac HC before being visualised AlphaImager software (ProteinSimple; California, USA) using the MultiImage Light Cabinet (Alpha Innotech Corporation; California, USA).

### *3.2.3.5 DNA clean up*

Fresh 80% ethanol was prepared by combining 40 ml of absolute ethanol with 10 ml UltraPure Distilled Water in a tornado tube. For each primer set, successfully amplified samples were pooled into a single 1.7 ml microcentrifuge. 25 µl of each sample were transferred into an Abgene storage plate (ThermoFisher; New Hampshire, USA) and 25 µl of AMPure XP beads (Agilent Technologies; California, USA) were

added. The plate was shaken at 1800 rpm for 2 minutes on the BioShake iQ (VWR Biotechnology; Pennsylvania, USA) and incubated at room temperature for 5 minutes. The plate was placed on a magnetic stand for ~2 minutes until the liquid was clear and the supernatant was discarded. The sample was washed twice with the 80% ethanol by adding 200 µl to each well, waiting 30 seconds, and then removing it. Excess ethanol was removed and the samples were allowed to air dry for ~5 minutes. The plate was removed from the magnetic stand and 25 µl UltraPure Distilled Water was added to each well. The plate was shaken at 1800 rpm for 2 minutes on the BioShake iQ and incubated at room temperature for 2 minutes. The plate was placed back on the magnetic stand until the liquid was clear. The supernatant was transferred to strip tubes and stored at 4ºC.

### *3.2.3.6 DNA quantification and normalisation*

Qubit (Invitrogen; California, USA) working solution was made by mixing the Qubit broad range reagent 1:200 in Qubit buffer, preparing 200 µl of working solution per sample, plus 200 µl for each of the two high sensitivity standards. 199 µl of working solution were added to Qubit assay tubes with 1 µl of sample per tube. For the standards, 190 µl of working solution were added to each tube with 10 µl of standard from the Qubit broad range kit. All tubes were vortexed for ~3 seconds each and allowed to stand at room temperature for 2 minutes. The concentrations in the tubes were read on the Qubit instrument as per

the manufacturer's instructions. 5 µl of each sample was diluted with UltraPure Distilled Water to a concentration of 0.2 ng/µl.

### *3.2.3.7 Tagmentation*

For this section, all reagents came from the Illumina (California, USA) Nextera XT kit unless stated otherwise. 10 µl of Tagment DNA buffer and 5 µl of normalised DNA were added to strip tubes, with each addition the pipette was used to mix. 5 µl Amplicon tagment mix was added to each tube and this was mixed with the pipette. The tubes were then placed in a centrifuge (VWR Galaxy MiniStar) at 2000 x g for ~20 seconds. The tubes were sealed with dome caps and placed on the GeneAmp PCR System 9700 thermocycler (Applied Biosystems; California, USA) at 55ºC for 5 minutes, and hold at 10ºC. 5 µl Neutralise Tagment Buffer was added to each tube and the tubes were centrifuged as described previously. The samples were then incubated at room temperature for 5 minutes.

### *3.2.3.8 Library indexing*

For this section, all reagents came from the Illumina Nextera XT library prep kit unless stated otherwise. Illumina Experiment Manager software was used to create a sample plate, giving each sample a unique index based on its row and column. To index and amplify the library, first 5 µl of index1 (i7) adapters were added to the columns of a PCR plate, one index per column. Next, 5 µl of the index 2 (i5) adapters were added to the rows of the PCR plate, one index per row. This provided each well

with a unique identifier based on which pair of indexes were in each well. 15 µl of Nextera PCR Master Mix were added to each well. For each sample, the entire sample from the tagmentation step was added to one of the wells with the well positions corresponding to the positions of the samples on the sample plate from Illumina experiment manager. The plate was sealed with Adhesive BioRad Microseal 'B' PCR plate seals. The plate was centrifuged on the Phenix (North Carolina, USA) Microplate Centrifuge for 1 minute and was then placed in the GeneAmp PCR System 9700 thermocycler (Applied Biosystems; California, USA) and run using the following conditions: 72ºC for 3 minutes. 12 x (95ºC for 10 seconds, 55ºC for 30 seconds, 72ºC for 30 seconds), 72ºC for 5 minutes. hold at 10ºC. DNA clean up was then performed as described previously in section 3.2.3.5 with the following alterations: the full 50 µl from the PCR plate were used for each sample, 30 µl of beads were used, the DNA was resuspended in 50 µl UltraPure Distilled Water.

1% agarose gel was prepared by mixing 6g agarose (VWR Biotechnology), 60 ml 5 x TBE and 540 ml pure $H_20$, swirling the bottle to mix, and microwaving for around 10 minutes on 50% power, stopping to swirl again at 5 minutes. The mixture was allowed to cool slightly (~5 minutes), then 16 µl Ethidium Bromide was added and the mixture was swirled again. A medium sized gel plate was used to run gel electrophoresis on the samples using the 1% agarose gel in a BioRad Wide Mini-Sub Cell. The gel was run at 70V for 1 hour and 30 minutes

on the BioRad PowerPac Basic before being visualised on AlphaImager software using the MultiImage Light Cabinet (Alpha Innotech Corporation).

### *3.2.3.9 Library denaturation, PhiX spike and loading the Illumina Miseq*

The Qubit was used as described previously in section 3.2.3.6 to quantify the concentration of DNA in each tube. The samples were then diluted to 4nM. 5 µl of each diluted sample was then pooled in a library in a microcentrifuge tube.

0.2 N NaOH was prepared by combining 800 µl UltraPure Distilled Water with 200 µl of 1.0 N NaOH, mixing by inversion. Illumina hybridisation buffer, HT1, was removed from the freezer and thawed and stored in a cool water bath until use. To denature the library, 5 µl of the library was combined with 5 µl of 0.2N NaOH. This was vortexed briefly and spun down in a centrifuge (VWR Galaxy MiniStar) for 30 seconds. The library was left at room temperature for 5 minutes, then 990 µl of HT1 was added, resulting in 1 ml of 20pM denatured library. The library was then diluted to 12pM by combining 360 µl of 20pM library with 240 µl HT1 in a new microcentrifuge tube. Stock 20pM denatured PhiX was diluted to 12.5pM by combining 75 µl PhiX with 45 ml HT1. In a new microcentrifuge tube, 594 µl denatured library was mixed with 6 µl 12.5pM denatured PhiX resulting in a 1% spike of PhiX as a control.

An Illumina sample sheet was created by following manufacturer's instructions and named using the barcode of the reagent cartridge being used for the sequencing run.

For loading the Miseq, the Illumina MiSeq Reagent Nano Kit v2 (300 cycle) was used. The reagent cartridge was thawed in a room temperature water bath for just over an hour. The reagent cartridge was gently inverted 10 times. A pipette tip was used to pierce the foil over the sample loading well. 600 µl of sample was pipetted into the sample loading well. The flow cell was rinsed with laboratory-grade water and gently dried with a tissue. The flow cell was loaded into the flow cell compartment of the Miseq as per manufacturer's instructions. Incorporation buffer (PR2) and the reagent cartridge were placed in the reagent compartment as per manufacturer's instructions and the waste bottle was emptied. The sequencing run was started as per manufacturer's instructions.

### *3.2.3.10 Bioinformatics*

Unless stated otherwise tools were used with default parameters. The fastq files produced by the MiSeq run were aligned to Sscrofa10.2 using BWA mem (v.0.7.15; Li, 2013) and output bam files were sorted and indexed using samtools sort and index (v. 1.2; Li et al., 2009a). Variants were called using Freebayes (v. 1.01; Garrison and Marth, 2012), the variant caller that was at the time being used as standard by the

industrial partner where this work was carried out, and visualised with IGV (v2.3.88) to confirm presence of variants.

### 3.2.4 Follow-up sequencing of sows

For this section the author selected the sows to sample, extracted and amplified the DNA and designed the primers, but the PCR and sequencing of the selected samples was carried out by Genus ABS employee Jennifer Dzelil. Bioinformatic analysis was carried out by the author.

Following validation of the variants, a subset of variants was selected as the focus of further investigation largely due to restricted time and resources. The selection of these variants was done through the methods described by Howard et al. (2017) to identify haplotypes in a different set of Genus pigs that, when present in sire and maternal grand sire, were significantly associated with litter size. Five of the prioritised variants were found by David Howard's work to occur within significant haplotypes. The five variants were in the genes ENSSSCG00000007065 (RPLP1), EXOG, LRRFIP2, SERPINA3 and LPIN3. Using Genus PIC's database, the litters of all of the boars carrying these variants were identified and information on litter size, number stillborn, number of mummies, identity of dam, among other details were obtained for all litters. For each variant, the 10 litters with the highest number of dead piglets were identified excluding those with a live litter size >15 (dead piglets may be due to lack of resources in a larger litter). The dams of these litters were identified and where

available tissue samples were retrieved from storage. All of the variants had at least one dam from the 10 litters selected for each variant with tissue available for sequencing, however tissue sample availability was severely limited. DNA was extracted from the tissue of 27 sows and quantified using the method described above for the boars. The 27 sows were from 5 different lines, only three individuals were from the same line as the boars. If one of the five variants is directly responsible for the increase in dead piglets, and a heterozygous cross can be found, the litter associated with this cross may be smaller or have a higher number stillborn than average.  Unfortunately, due to a relocation of the Genus ABS/PIC lab, the tissue samples have since been lost and this DNA is all that remains from these sows.

#### *3.2.4.1 Whole genome amplification*

In order to conserve DNA from these sows, a whole genome amplification was carried out using the REPLI-g Mini Kit (Qiagen;Germany). While unamplified DNA is preferred, this DNA served as a backup. Buffer D1 and buffer N1 were prepared according to manufacturer's instructions, all reagents were vortexed and spun down briefly in a centrifuge prior to use. DNA was diluted to 40 ng/µl prior to use using UltraPure Distilled Water. 2.5 µl of each sample was combined with 2.5 µl Buffer D1, mixed by gentle pipetting and incubated at RT for 3 minutes. 5 µl Buffer N1 was added to the samples and mixed by gentle pipetting to neutralise the reaction. A master mix was created using the following volumes for each sample: 10 µl UltraPure

Distilled Water, 29 µl reaction buffer and 1 µl DNA polymerase. 40 µl of the mix was added to each sample

### *3.2.4.2 Primer design and resuspension*

For this second sequencing run, the same set of primers were used as the validation sequencing with the exception of the variants that failed to sequence or validate previously, for which new primers were designed as described previously, these were the primers for variants in ENSSSCG00000004427 (NPM1-like), TRRAP and FAM216B. Additional primers were also designed for the VCAN variant, which failed to amplify on some of the samples using the previously designed primer pair. Although five variants were the main focus for these mating pairs, primers were ordered for all 20 regions of interest. The primers were designed with Illumina adaptors. These primers were designed, ordered and resuspended as described previously.

### *3.2.4.3 PCR*

PCR was carried out using 12.5 µl NEB 2x taq, 1 µl forward primer (25 µM), 1 µl reverse primer (25 µM), 2 µl DNA and 8.5 µl UltraPure Distilled Water per primer pair, per individual. PCR was run on the MultiGene OptiMax Thermal Cycler TC9610 (Thermo Fischer) for 94ºC for 4 minutes, 30 x (95ºC for 30 seconds, 54ºC for 30 seconds, 72ºC for 90 seconds) and 72ºC for 5 minutes, holding at 4ºC. For most primer pairs, WGA DNA was used, but for the primer pairs for variants in LAMA4, MTHFD1, MST1R, ENSSSCG00000017574 (RDM1), LAMB3,

ENSSSCG00000004427 (NPM1-like) and FAM216B, the unamplified DNA was available and was used.

### *3.2.4.4 Library indexing and clean up*

Amplicons from each individual were pooled into a single sample and were normalised based on quantification results from Qubit, which was carried out as described previously. Nextera XT Library Prep Kit was used according to protocol to create dual-indexed libraries, reagents are from the Nextera XT library prep kit unless stated otherwise. 10 µl of Tagment DNA Buffer (TD) was added to wells of a hard-shell skirted PCR plate. 5 µl of normalized sample was added to each well and mixed via pipetting. 5 µl of Amplicon Tagment Mix (ATM) was added to each well and mixed via pipetting. Samples were incubated at 55°C for 5 minutes on MultiGene OptiMax Thermal Cycler TC9610 (Labnet International) and then held at 10°C. Once the sample was cooled to 10°C, 5 µl of Neutralize Tagment Buffer (NT) was added to each well, pipetting up and down to mix, and incubated at room temperature for 5 minutes. 5 µl of Index 1 (i7) adapter was added to each column of samples. 5 µl of Index 2 (i5) adapter was added across each row. 15 µl of Nextera PCR Master Mix (NPM) was added to each sample and mixed via pipetting. Samples were then amplified on MultiGene OptiMax Thermal Cycler TC9610 (Thermo Fischer) with the following conditions: 72°C for 3 minutes, 95°C for 30 seconds, 12 x (95°C for 10 seconds, 55°C for 30 seconds, 72°C for 30 seconds), 72°C for 5

minutes, 4°C hold. A gel electrophoresis was run to check for the presence of amplicons.

A bead clean-up was performed as follows to try to reduce the dimer in the sample. For each pool, 50 µl was taken and placed in a column of a deep well midi plate and 60 µl in-house magnetic beads (as described by Rohland and Reich, 2012) were added to each well. The plate was shaken at 1800 rpm for 2 minutes on the BioShake iQ and incubated at room temperature for 5 minutes. The plate was placed on a magnet, the beads were allowed to pellet, and the supernatant was discarded. 200 µl fresh 80% ethanol was added to each well without disturbing the pellet, left for 30 seconds, and was discarded, this was done twice. Excess ethanol was removed and the plate was left at room temperature for 5 minutes before being removed from the magnet. 25 µl Resuspension Buffer (RSB) was added to each well, the plate was shaken at 1800rpm for 2 minutes and incubated at room temperature for 2 minutes. The plate was placed back on the magnet for 2 minutes and the eluted samples were transferred to a fresh plate. All samples were pooled into one tube and quantified using Qubit as described previously. The library was diluted to 4nM using UltraPure Distilled Water.

### *3.2.4.5 Library denaturation, PhiX spike and loading the Illumina Miseq*

Reagents are from the Illumina Miseq v3 kit unless stated otherwise. To denature the library, 5 µl was transferred to a new tube and 5 µl 0.2 N NaOH was added. The tube was vortexed and briefly centrifuged before being incubated at room temperature for 5 minutes and 990 µl chilled HT1 was added resulting in a 20 pM denatured library. The library was diluted to 10 pM by combining 300 µl library with 300 µl chilled HT1.

PhiX was diluted to 4 nM by combining 2 µl 10nM PhiX with 20 mM Tris-Cl, pH 8.5 with 3 µl 0.1% Tween 20. 5 µl of 4 nM PhiX was combined with 5 µl 0.2N NaOH and briefly vortexed and centrifuged. This was incubated at room temperature for 5 minutes. The denatured PhiX was diluted to 20 pM by combining 10 µl with 990 µl of chilled HT1. The PhiX was diluted to 10 pM by combining 300 µl with 300 µl of HT1. The libraries were combined with the PhiX to make a 600 µl final library, the first library had a 25% PhiX spike in, and the second library had a 15% PhiX spike in.

A sample sheet was created by following manufacturer's instructions and named using the barcode of the reagent cartridge being used for the sequencing run.

For loading the MiSeq, the Illumina MiSeq Reagent Nano Kit v3 (600 cycle) was used. The reagent cartridge was thawed in a room temperature water bath for just over an hour. The reagent cartridge was

gently inverted 10 times. A pipette tip was used to pierce the foil over the sample loading well. 600 µl of sample was pipetted into the sample loading well. The flow cell was rinsed with laboratory-grade water and gently dried with a Kimtech tissue (Kimberly-Clark; Texas, USA). The flow cell was loaded into the flow cell compartment of the MiSeq as per manufacturer's instructions. Incorporation buffer (PR2) and the reagent cartridge were placed in the reagent compartment as per manufacturer's instructions and the waste bottle was emptied. The sequencing run was started as per manufacturer's instructions.

#### *3.2.4.6 Bioinformatics*

Unless stated otherwise, tools were run using default parameters. The fastq files produced by the MiSeq run were aligned to Sscrofa10.2 using BWA mem (v.0.7.15; Li, 2013) and bam files were sorted and indexed using samtools sort and index (v1.2; Li et al., 2009a). Variants were called using Platypus (v. 0.8.1; Rimmer et al., 2014), at the time the variant caller being used by the industrial partner where this work was carried out, and visualised with IGV (v2.3.88; Thorvaldsdóttir et al., 2013) to confirm presence of variants. A Fisher’s exact test was used to identify significant differences in the presence of alleles between males and females, however these differences may in part reflect differences in breeding lines.

### 3.2.5 Imputation and phenotype association

#### *3.2.5.1 Data*

As the two available phenotypes are only subtly different, the EBVs were plotted against one another to check for any correlation.

Available EBVs were filtered to keep only those individuals with a reliability score >0.5. While the higher reliability EBVs have the most accurate phenotyping, there are a reduced number of individuals with phenotypes for the association analysis which will make the association analysis less accurate. The filtered dataset has EBVs for only 204 individuals for NSB and 224 individuals for BTP.

#### *3.2.5.2 Imputation*

The 96 exome pigs were used as a reference population, the variants from exome sequencing (excluding variants from the LQLC regions) were used as the more high density genotype data set and Beagle (v 3.3.2; Browning and Browning, 2009) was used to impute genotypes for the test population from the PorcineSNP60 BeadChip (Illumina) to the exome set.

#### *3.2.5.3 Phenotype association*

Plink (v 1.07; Purcell et al., 2007) was used with default parameters to run a whole genome association analysis of the alleles against the two EBVs. Power was estimated for these GWAS using R (v.3.3.3; R Development Core Team, 2009) using the following commands:

N = 204 (or 224 for BTP)
alpha = 0.00000025
H2 = 0.04 (or 0.07 for BTP)

threshold = qchisq(alpha, df = 1, lower.tail = FALSE)
power = pchisq(threshold, df = 1, lower.tail = FALSE, ncp = N * R2)

This method is described at https://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits (Accessed 18/04/19). Where N is the number of individuals, alpha is 0.5/the number of markers and H2 is the heritability of the trait as estimated by Holm et al. (2004). The power was found to be 0.0107 for NSB and 0.1154 for BTP. This makes the power of these analyses extremely low, unfortunately no further data was available to increase the power. The power was also estimated for a range of Ns to demonstrate how many additional individuals would be needed to increase the power for future experiments for BTP, as an example.

### *3.2.5.4 Candidate variant selection*

A cut off for p-values was set based on the Bonferroni correction of 0.05/N. For each phenotype, a bed file was made of 5Kb either side of the coordinates of the significant variants and the regions were merged using BEDtools merge. From the original VCF of variants from the exome data prior to LQLC filtering, variants that overlapped these regions were extracted using BEDtools intersect. Local misassemblies of the genomes in significant regions may have led to causative variants being filtered out in the LQLC filter, or may disrupt accurate

identification of haplotypes during imputation. Variants in these regions that were predicted to have a moderate or high impact on the protein were identified. A literature search was carried out for named genes containing these variants to find any functional association with reproductive traits or embryonic survival.

#### *3.2.5.5 Candidate gene selection*

Regions of the genome that showed a strong association were searched for candidate genes. An associated region may be linked to a regulatory element or splice site disruption that influences the expression of a gene, however these are likely outside of the exome region and further investigation of these is outside the scope of this project. Additionally, variants from the exome data were subsequently filtered out of the dataset to see if the same candidates are identified without the additional data from exome sequencing.

## 3.3 Results

### 3.3.1 Candidates based on filtering without phenotype data

#### *3.3.1.1 Filtering out LQLC regions*

26.37% of the coding region was classified as LQLC as defined in chapter 2. Removing variants found in the LQLC regions reduced the number of SNPs from 236,530 to 150,600 and the number of indels from 28,976 to 12,544, a reduction of 36.33% and 43.3%, respectively. The Ts/Tv ratio of the SNPs prior to this filter was 2.82, following filtering the Ts/Tv ratio increased to 2.92, suggesting an increase in

accuracy assuming the expected value of 3 in the protein coding regions. The variants in the LQLC regions had a Ts/Tv of 2.64, further suggesting this set contains a higher proportion of errors compared to the unfiltered set.

#### *3.3.1.2 Variant annotation and consequence filtering*

Table 3-1 shows the number of variants in each consequence category for SNPs and indels in the original data set, the data set after filtering out LQLC regions, and after filtering for variants in the first half of the exons and annotated as having a high impact.

Following filtering for variants where at least 70% of the carriers have at least 40X coverage there are 145 high impact SNPs and 310 high impact indels.

**Table 3-1- Table showing the breakdown of variant types in the 96 exomes as annotated by Ensembl VEP including raw data, data filtered for LQLC regions, and data filtered for LQLC, position in the exon, and predicted impact.**

| | SNP | SNP-LQLC | SNP-LQLC, Exon<=0.5, HIGH impact | Indel | Indel - LQLC | Indel-LQLC, Exon<=0.5, HIGH impact |
|---|---|---|---|---|---|---|
| **Stop gain** | 720 | 411 | 180 | 33 | 10 | 4 |
| **Start loss** | 134 | 71 | 1 | 35 | 8 | 8 |
| **Frameshift** | - | - | - | 4,292 | 958 | 546 |
| **Missense** | 41,527 | 22,923 | - | - | - | - |
| **Inframe indel** | - | - | - | 784 | 375 | - |
| **Stop lost** | 283 | 144 | - | 38 | 6 | - |
| **Splice** | 2,822 | 1,545 | - | 817 | 125 | - |
| **Synonymous** | 64,991 | 38,076 | - | - | - | - |
| **Total** | 236,530 | 150,600 | 180 | 28,976 | 16,431 | 547 |

**Table 3-2- Indels that are not in Hardy-Weinberg Equilibrium and are in named genes. Parentheses are used to differentiate between multiple variants identified in the same gene.**

| Gene | Expected | Observed |
|---|---|---|
| SUB1(1) | 38.13/44.74/13.13 | 25/71/0 |
| SUB1(2) | 49.59/38.81/7.59 | 42/54/0 |
| DSTN | 27.63/47.74/20.63 | 7/89/0 |
| CALM1 | 25.01/47.98/23.01 | 2/94/0 |
| DTNB(1) | 24.00/48.00/24.00 | 0/96/0 |
| DTNB(2) | 34.44/46.12/15.44 | 19/77/0 |
| CCDC30 | 25.01/47.98/23.01 | 2/94/0 |

### *3.3.1.3 Prioritising based on Hardy-Weinberg Equilibrium*

Of the high consequence variants in the first 50% of the total exons of the gene, with 0 homozygotes, 7 SNPs and 29 indels were not in HWE. Of the 7 SNPs, two are in known olfactory receptors and two are in loci that show sequence similarity to olfactory receptors in other species. Of the remaining SNPs, one is in a loci with sequence showing similarity to a ribosomal protein gene, one is in an unnamed gene and has no clear orthologue and the final one is within gene SERPINA3, which has an expected genotype distribution of 58.59/32.81/4.59 and an observed distribution of 54/42/0 with a p value of 0.00519 (exact test). Of the 29 indels, five are in known olfactory receptors and 12 show sequence similarity to olfactory receptors in other species. Of the remaining indels, four are in unnamed genes with no clear orthologues and the remainder have a P value <0.00001 are listed in table 3-2. No variants were excluded based on HWE.

### *3.3.1.4 Filtering and prioritising based on gene function*

After removal of genes related to olfactory functions, 77 SNPs and 128 indels remain. Priority was given to named genes in either the Ensembl or NCBI annotation of Sscrofa10.2, and those with orthologues identified through the Ensembl website, for which no homozygous individuals were found in the sample population. These include 27 SNPs affecting 27 genes and 38 indels affecting 35 genes. From this list several variants were excluded due to 6-frame translations or visualisation of the alignments disagreed with Ensembl VEP. An

example of variants excluded during this process are frameshifts in the gene KDM5A, where two frameshifts 4 bases apart were consistently called on the same haplotype and in 6-frame translations the second frameshift rescued the protein by shifting the translation back into the correct frame. The SNP in SERPINA3 is the only variant that was significant for deviation from HWE and passed these filters. From these restrictions variants were prioritised and are listed in table 3-3. These variants, along with those discussed previously in CFTR (x2), IL1RAP and TGFB2 were taken forward to the validation stage.

### 3.3.2 Validation of prioritised variants

19 of the prioritised variants were present in amplicons from resequencing the regions in the boars, with three failing to validate. One of those that failed to validate, the indel in ENSSSCG00000004427, did not successfully amplify, suggesting the primers failed.

The other two variants in TRRAP and FAM216B were not seen in the alignments of the sequenced amplicons. In the case of the indel in TRRAP, a nearby SNP on the same haplotype in the exome data was also absent. Two individuals were sequenced for this variant, but it is possible that through human error the wrong individuals were sequenced. In the case of the SNP in FAM216B, There are two neighbouring SNPs that in some exome alignments appear to be on the same haplotype as the variant, and in others appear to be independent. Two individuals were sequenced here and while the neighbouring SNPs

appear heterozygous, the SNP of interest is absent. It is not clear why these two variants failed to validate.

**Table 3-3- Table listing the genes containing variants that were followed up on following filtering. Hom=homozygous reference, Het=heterozygous. Only the last three rows have any homozygous alt individuals (not shown).**

| Gene symbol | Gene name | Variant type | Variant consequence | Hom | Het |
|---|---|---|---|---|---|
| LRRFIP2 | LRR Binding FLII Interacting Protein 2 | SNP | Stop gain | 80 | 16 |
| LAMA4 | Laminin Subunit Alpha 4 | SNP | Stop gain | 84 | 12 |
| SERPINA3 | Serpin Family A Member 3 | SNP | Stop gain | 54 | 42 |
| WDR36 | WD Repeat Domain 36 | SNP | Stop gain | 83 | 13 |
| MST1R | Macrophage Stimulating 1 Receptor | SNP | Stop gain | 94 | 2 |
| LPIN3 | Lipin 3 | SNP | Stop gain | 95 | 1 |
| ENSSSCG00000017574 (fragment of RDM1) | (RAD52 Motif Containing 1) | SNP | Stop gain | 90 | 6 |
| LAMB3 | Laminin Subunit Beta 3 | SNP | Stop gain | 93 | 3 |
| EXOG | Exo/Endonuclease G | SNP | Stop gain | 94 | 2 |
| FAM216B | Family With Sequence Similarity 216 Member B | SNP | Stop gain | 92 | 4 |
| ENSSSCG00000004427 (NPM1-like gene) | (Nucleophosmin) | Indel | Frameshift | 82 | 14 |
| TPT1 | Tumor Protein, Translationally-Controlled 1 | Indel | Frameshift | 89 | 7 |
| ENSSSCG00000007065 (fragment of RPLP1) | (Ribosomal Protein Lateral Stalk Subunit P1) | Indel | Frameshift | 95 | 1 |
| ENSSSCG00000012880 (fragment of CPT1A) | (Carnitine Palmitoyltransferase 1A) | Indel | Frameshift | 90 | 6 |
| VCAN | Versican | Indel | Frameshift | 95 | 1 |
| TRRAP | Transformation/Transcription Domain Associated Protein | Indel | Frameshift | 91 | 5 |
| MTHFD1 | Methylenetetrahydrofolate Dehydrogenase, Cyclohydrolase And Formyltetrahydrofolate Synthetase 1 | Indel | Frameshift | 95 | 1 |
| CFTR | Cystic fibrosis | 2 x | 2 x missense | 87 | 9 |

| | | | | | |
|---|---|---|---|---|---|
| | transmembrane conductance regulator | SNP | | 41 | 42 |
| IL1RAP | Interleukin 1 Receptor Accessory Protein | SNP | Missense | 83 | 11 |
| TGFB2 | Transforming Growth Factor Beta 2 | SNP | Missense | 58 | 31 |

### 3.3.3 Selection and follow up sequencing of sows

The average litter size (dead + alive) of all of the litters sired by all of the males carrying one of the five prioritised variants was 12.5, with an average percent dead of 9.57% (1.2 observed dead per litter). The average litter size (dead + alive) of the females selected for follow up sequencing was 13.6 with an average percent dead of 28.4% (3.9 observed dead per litter). The number dead per litter ranged from 1 to 9, and the percent dead per litter ranged from 7.7% to 81.8%. 54 dams were identified that were mated with males carrying the primary focus variants in LPIN3, ENSSSCG00000007065 (RPLP), EXOG, LRRFIP, or SERIPNA3 as described in 3.2.4, however only 27 of these had tissue available for sequencing and these did not equally represent all five of the variants. While dams from 27 litters were genotyped, these represent offspring from only 9 males as 2 of the males sired 10 of the litters each, with fewer tissue samples available for the dams of litters from other sires. Crosses between individuals that were carriers of the same putative LoF variants were observed for the variants in EXOG, LRRFIP2, SERPINA3 and LAMB3 and while all of these litters had average litter sizes above that of all of the litters sired by the boars at 14.5, 13.8, 13 and 13.6, respectively, they also had a higher than average percent of the litter dead at 14.5%, 28.7%, 40.6% and 21%, respectively. The highest average number of dead piglets per litter in a heterozygous cross was 5, which is around 4 times the average

**Table 3-4- Table showing difference in allele frequencies of exome candidate variants between males and females. P-values from Fisher's exact test for presence/absence of the allele between groups.**

| Gene containing variant | Frequency in females | Frequency in males | Difference | p-value |
|---|---|---|---|---|
| LRRFIP2 | 0.185 | 0.083 | +10.2% | 0.032 |
| LAMA4 | 0.093 | 0.063 | +3.0% | 0.527 |
| SERPINA3 | 0.241 | 0.219 | +2.2% | 0.827 |
| IL1RAP | 0.019 | 0.078 | -6.0% | 0.3 |
| MST1R | 0.000 | 0.010 | -1.0% | 1 |
| TGFB2 | 0.204 | 0.234 | -3.1% | 1 |
| RDM1 | 0.241 | 0.031 | +20.9% | 2.05E-06 |
| LPIN3 | 0.000 | 0.005 | -0.5% | 1 |
| EXOG | 0.111 | 0.010 | +10.1% | 0.0056 |
| FAM216B | 0.000 | 0.021 | -2.1% | 0.575 |
| NPM1 | 0.019 | 0.073 | -5.4% | 0.187 |
| TPT1 | 0.000 | 0.036 | -3.6% | 0.346 |
| RPLP1 | 0.093 | 0.005 | +8.7% | 0.033 |
| CPT1A | 0.463 | 0.031 | +43.2% | 2.75E-18 |
| VCAN | 0.000 | 0.005 | -0.5% | 1 |
| TRRAP | 0.111 | 0.026 | +8.5% | 0.014 |
| MTHFD1 | 0.130 | 0.005 | +12.4% | 0.00041 |
| WDR36 | 0.019 | 0.068 | -4.9% | 0.3 |
| LAMB3 | 0.185 | 0.016 | +17.0% | 0.00001 |
| CFTR(1) | 0.000 | 0.047 | -4.7% | 0.204 |
| CFTR(2) | 0.296 | 0.354 | -5.8% | 1 |

**Table 3-5- Table showing the litters where a dam was identified as being homozygous for a variant.**

| | EXOG | RPLP1 | RPLP1 | MTHFD1 | CFTR(2) |
|---|---|---|---|---|---|
| **Dam genotype** | 1/1 | 1/1 | 1/1 | 1/1 | 1/1 |
| **Sire genotype** | 0/1 | 0/0 | 0/0 | 0/0 | 0/0 |
| **Litter size** | 14 | 18 | 13 | 15 | 9 |
| **Number observed dead** | 5 | 3 | 2 | 4 | 7 |
| **Percentage dead** | 35.71% | 16.67% | 15.38% | 26.67% | 77.78% |

for all of the litters of the sires, and was associated with the variant in SERPINA3. No correlation was observed between parental genotypes of any of the variants and the litter size or percent dead of the litters in the 27 litters.

There were some notable differences in allele frequencies between the males and females, the most striking of which was the indel in ENSSSCG00000012880 (CPT1A), which was only seen 6 times in the exome data (allele frequency of 0.031), but was seen in 25 of the 27 females (allele frequency of 0.463) and had a Fisher's exact test p-value of 2.75E-18. Table 3-4 shows the differences in frequencies between the males and females for all of the variants. Four of the variants were seen in the homozygous form, these are shown in table 3-5. This suggests that these four variants are not lethal or else do not have full penetrance. However, the second missense variant in CFTR, which is very common in both the males and the females, is homozygous in only one of the females and is associated with one of the least successful litters with only 2 surviving offspring. This may suggest a maternal phenotype, however with only one example of a homozygous female it is impossible to say if this is related. Two of the variants that failed to validate in the exome pigs were found in the sows, these are the variants in TRRAP and ENSSSCG00000004427 (NPM1-like).

### 3.3.4 Candidates from imputation and phenotype association

Following imputation from SNP chip to exome, GWAS were carried out for the two EBVs number stillborn (NSB) and number born to purebred (BTP) to search for association between genomic loci and these traits. The power of this analysis was very low and further individuals were not made available for this thesis to increase the power. Figure 3-2 shows how the power for BTP would increase in the analysis with the addition of individuals.

There is no correlation between NSB and BTP (p=0.255; figure 3-3).

Only 4 variants reached significance for BTP. 5Kb either side of these 4 variants were merged resulting in three regions. Within these regions were 15 variants from the exome dataset. None of these variants were high or moderate impact variants. The genes within, or closest to, these regions include COL11A1, ZFYVE27, and SFRP5.

102 variants reached significance for NSB, merging of 5Kb either side resulted in 46 regions. Within the 46 regions were 241 variants from the exome data set of which there were three high impact variants and 23 moderate impact variants, 11 of which had no observed homozygotes. Two of the high impact variants were in unnamed genes, ENSSSCG00000004082 (SYNE1) and ENSSSCG00000026692 (likely ARHGEF10), and the third was in TSHZ1, however this third variant was homozygous in all individuals and is likely benign or a false-positive. Of the 11 missense SNPs with no observed homozygotes, 4

were not marked tolerated by SIFT. Of these variants, two were in unnamed genes and two were in named genes and both of these were marked deleterious by SIFT. These were in the genes CORO2B (SIFT score 0) and SERPINB8 (SIFT score 0.05). There were 5 variants that were not marked tolerated by SIFT, but did have homozygotes present. These were all in unnamed genes, one of which is likely MYOM2 (ENSSSCG00000015747) and is marked deleterious (0.02) by SIFT, while the others are in unidentified refseq genes.

Figure 3-4 shows a Manhattan plot for BTP. Two peaks reach genome-wide significance, however the plot is very noisy likely due to the limitations of the available data.

Figure 3-5 shows a Manhattan plot for the BTP data without the exome data included. A peak on chromosome 10 that only approached genome-wide significance when the exome data was included does reach significance with the lower p-value from the reduced number of tests when they are excluded.

Figure 3-6 shows a Manhattan plot for NSB. Several regions in the plot reach genome-wide significance and appear to be stronger peaks than those identified for BTP.

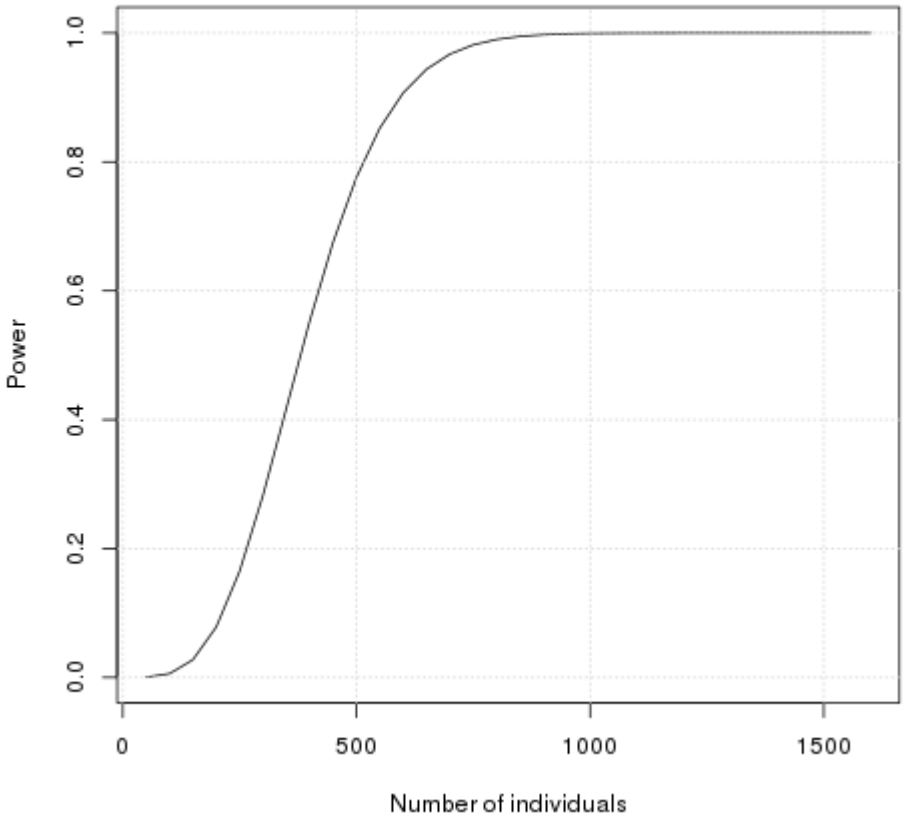



**Figure 3-2 Figure showing the effect on the estimated power of increasing the sample size for BTP.**

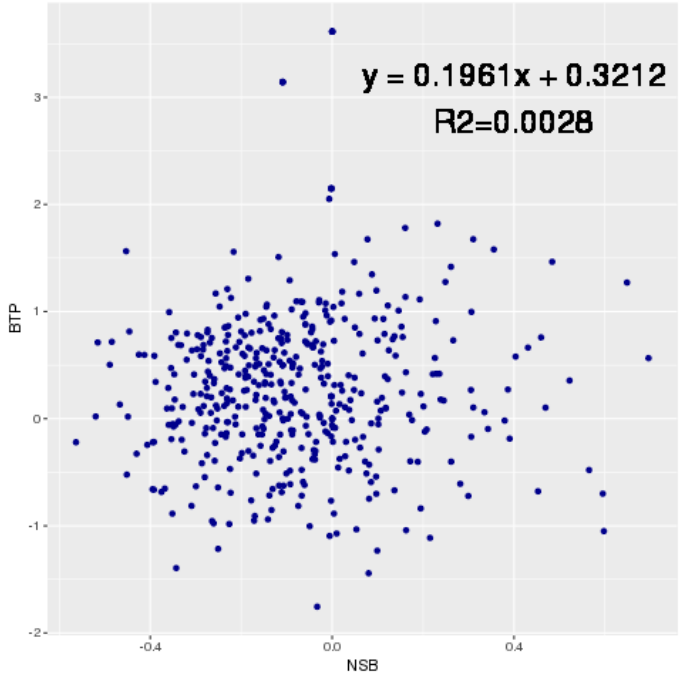



**Figure 3-3- Plot of NSB against BTP showing no correlation between the two phenotypes**

A region containing a significant variant at 15:138464069 next to an assembly gap is a ~21 Kb contig containing fragments of SFXN5 and RAB11FIP5, both of which also appear fragmented on chromosome 3. The contig on chromosome 15 is misplaced and should be located on chromosome 3 at ~73 Mb, where there is a peak approaching significance over SFXN5. The placement of the contig on chromosome 3 is supported by the gene order observed in the genomes of other species.

Figure 3-7 shows the same NSB data without the exome data included. In this case most of the significant regions in figure 3-6 are also present here, with a gain of a significant variant on chromosomes 3 and 10.

Table 3-6 describes the genes found under the peaks in figures 3-4 to 3-7 and the location of the most significant variant for each of these peaks.

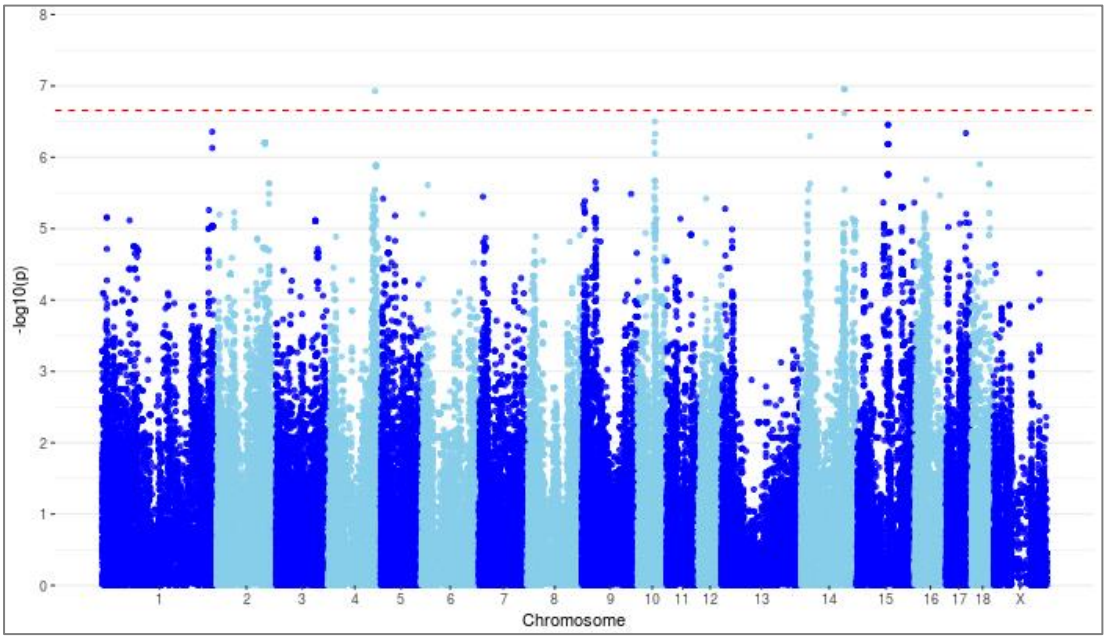



**Figure 3-4- Manhattan plot showing association between exome and SNP chip variants and BTP. Red dashed line shows genome-wide significance level.**

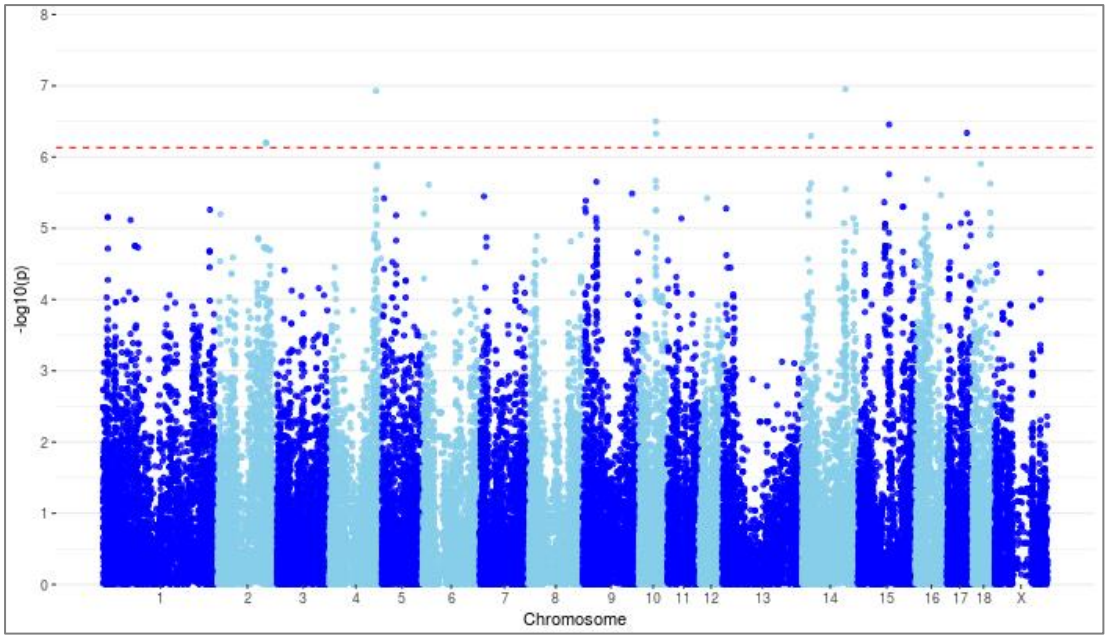



**Figure 3-5- Manhattan plot showing association between SNP chip variants and BTP. Red dashed line shows genome-wide significance level.**

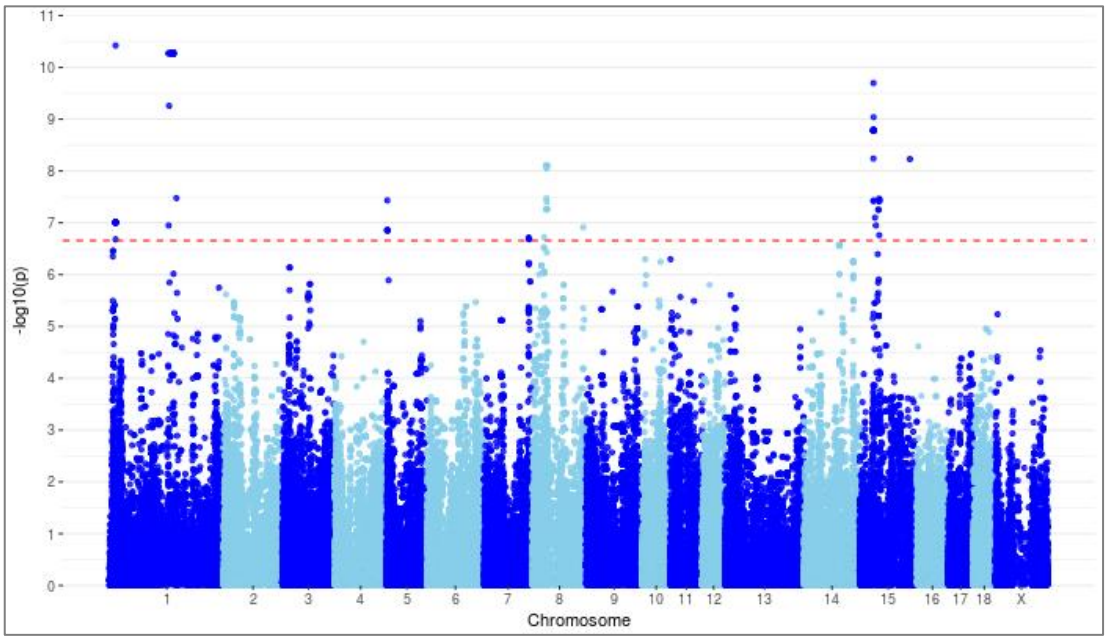



**Figure 3-6- Manhattan plot showing association between exome and SNP chip variants and NSB. Red dashed line shows genome-wide significance level.**

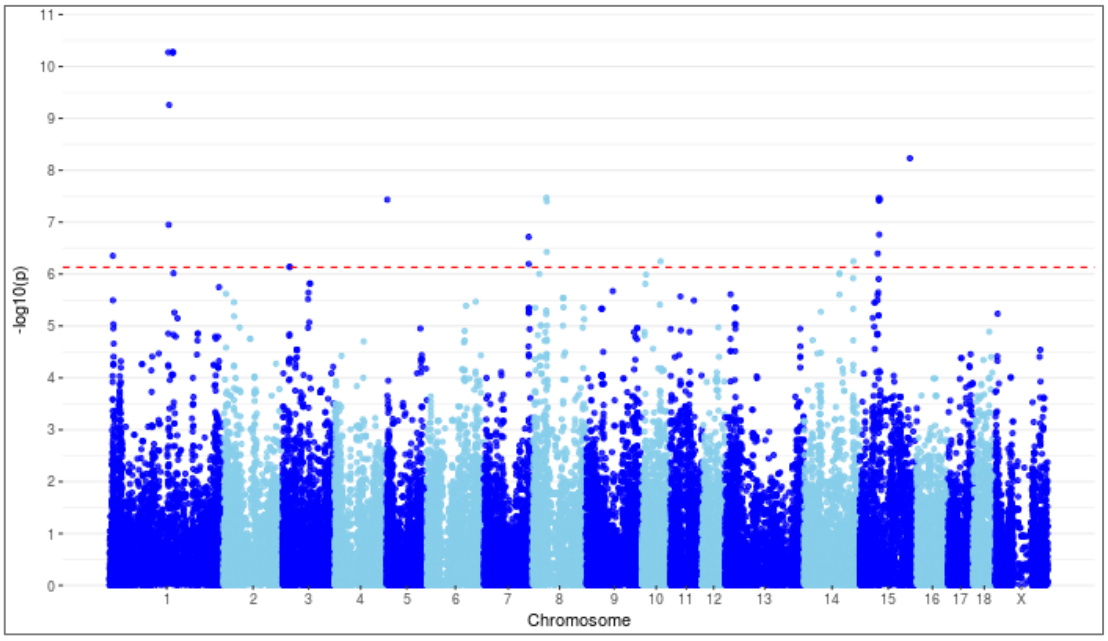



**Figure 3-7- Manhattan plot showing association between SNP chip variants and NSB. Red dashed line shows genome-wide significance level.**

**Table 3-6- Table showing the location of the most significant variant and genes located under peaks for each peak in figures 3-4 to 3-7. Regions marked "without exome" are based on SNP chip variants alone.**

| EBV | With/without exome | Position of most significant variant | Genes near significant variants |
|---|---|---|---|
| BTP | With exome | 4: 127489069 | COL11A1 |
| | | 14: 118637605 | ZFYVE27, SFRP5 |
| | Without exome | 10: 47043706 | ARHGAP12, ENSSSCG00000011025 (ZEB1) |
| | | 2: 132097143 | CEP120, PRDM6 |
| | | 14: 23840807 | ZNF268, ZNF84 |
| | | 15: 84843754 | DHRS9 |
| | | 17: 54169497 | ELMO2, SLCR2 |
| NSB | With exome | 1: 16507269 | ENSSSCG00000004082, ESR1 |
| | | 1:162345314-175822562 (32 variants in LD) | SERPINB cluster |
| | | 5: 2157220 | PARVB |
| | | 7: 123675137 | SERPINA cluster, TCL1A, TCL1B, GLRX5, SYNE3, DICER1, CLMN, BDKRB2, BDKRB1, ATG2B |
| | | 8: 38744361-38745803 (8 variants in LD) | GABRG1, GABRA2, GABRA4 |
| | | 8: 139878201 | HERC3 |
| | | 15: 37979158-38266589 (13 variants in LD) | ENSSSCG00000015747 (MYOM2 orthologue), ENSSSCG00000023419 (ARHGEF10 orthologue) |
| | | 15: 138464069 | ENSSSCG00000024710 (SFXN5 fragment), ENSSSCG00000024710 (RAB11FIP5 fragment) |
| | Without exome | 3: 20059846 | ENSSSCG00000007821 (Pseudogenes) |

Many of the most significant variants were in the exome data, and the loss of these data results in different or less specific regions being highlighted. For example, in the major peak on chromosome 15, the most significant variant in the SNP chip data occurs at 15: 53557174, whereas in the exome data the most significant variants cluster over 2 genes, ENSSSCG00000015747 (MYOM2 orthologue) and ENSSSCG00000023419 (ARHGEF10 orthologue), including the most significant variant at 15: 37993862.

On chromosome 8, the SNP chip variants highlight a region of over 1 Mb as shown in the Ensembl Genome Browser (Zerbino et al., 2018) in figure 3-8 with the most significant of the three at 8: 37838817.

The exome variants highlight a smaller region of 2.5Kb (figure 3-9), with the majority of the significant variants falling within GABRA4 and the most significant variant at 8: 38745803.

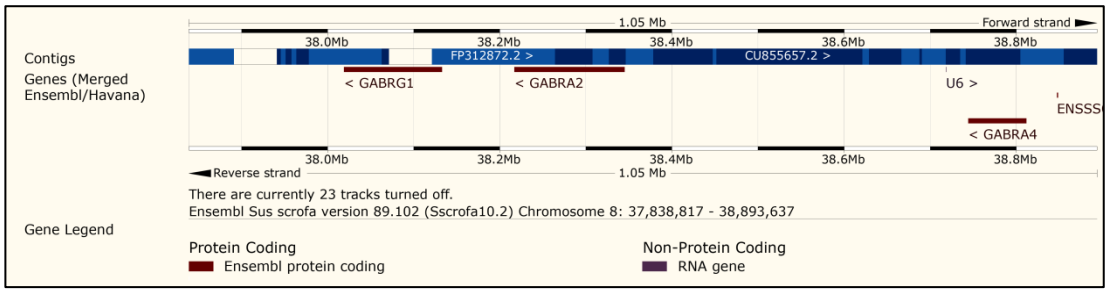



**Figure 3-8- Ensembl genome browser visualisation of 1Mb region on chromosome 8 containing SNP chip variants associated with NSB.**

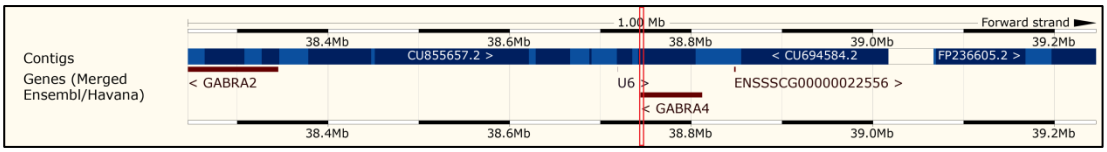



**Figure 3-9- Ensembl genome browser visualisation of 2.5 Kb region on chromosome 8 containing exome variants associated with NSB marked in red box.**

## 3.4 Discussion

Reproductive traits are an important target for selection in the pig. Detrimental variants in the nucleus herds of pig production companies can have a knock on effect on the productivity of the rest of the breeding pyramid. Identification of variants that effect these traits can allow for purging of detrimental variants. Reproductive phenotypes are complex and careful filtering combined with an understanding of gene functions and identification of associations with phenotypes can aid in identifying variants of interest.

### 3.4.1 Exome filtering

For the pig genome, there is relatively little variant calling data available to assist with variant quality control and filtering and the data that exists may not be reliable. Over 50% of Sscrofa10.2 variants reported in dbSNP (Sherry et al., 2001) in unreliable regions of the genome and likely enriched with false-positives (Warr et al., 2015b). One method for assessing the accuracy of variant calling data is to look at the ratio between transitions and transversions, which in humans has been reported as 3:1 in the coding regions (Heinrich et al., 2013, Zhang et al., 2015, Bainbridge et al., 2011). A transition is a substitution of either a purine (A,G) with the other purine or pyrimidine (T,C) with the other pyrimidine, and a transversion is a substitution of a purine to a pyrimidine or vice-versa. Transitions are more common than transversions due to the mechanisms through which the substitutions are formed. Transitions are even more common than transversions in

the coding region as they are less likely to change the amino acid thanks to the redundancy in codon translations, often creating a synonymous variant and avoiding the negative selection that limits transversions (Griffiths et al., 2015). In the pig, the Ts/Tv ratio for the whole genome has been reported to be around 2.04:1 (Bianco et al., 2015) which is similar to the 2:1 ratio seen in humans (Bainbridge et al., 2011). As the data in this project deals with exome sequencing data, the Ts/Tv would be expected to be closer to 3:1 as it is in the coding region in humans (Bainbridge et al., 2011). In this project, the original data set had a Ts/Tv of 2.82:1, the exonic variants that fall within the LQLC region previously identified have a Ts/Tv of 2.64:1, and removing these variants increased the Ts/Tv of the dataset to 2.92:1, suggesting this filter reduced the false positive rate. The LQLC region covered 26.37% of the coding region and eliminated 36.33% of the SNPs and 43.3% of the indels. While many of these eliminated variants may have been true positives, the elevated proportion of variants located in these regions further suggests that they were enriched for false-positives.

While the LQLC filter greatly reduced the number of variants, a combined total of over 170,000 remained. Annotating the variants allows for elimination of variants that are the least likely to have a direct impact on the protein, and prioritise those that may truncate the protein. By selecting for variants that occur early in the gene, the chance of a variant knocking out the functional portion of the protein is increased. Additionally, the confidence that the variant is not a false-positive call is

increased by prioritising variants that have at least 40X coverage in at least 70% of the individuals that carry it. These filters reduced the number of variants to a more manageable 455 variants.

HWE is a model for predicting expected genotype frequencies in a population given the frequencies of the two alleles. Deviation from the equilibrium may suggest genotyping error, a copy number variation, or natural or artificial strong selective pressures for or against a haplotype. While the HWE filter did find one SNP that may be interesting to follow up on, in SERPINA3, many of the variants identified had very unlikely genotype frequencies, for example, those variants where almost all individuals are heterozygous. In these cases of heterozygote excess, it is possible there is some kind of duplicated region or paralog, which may not be represented in the genome assembly, causing the site to look like a heterozygous SNP when in fact there is no variant but rather sequences from two regions aligning to one (Wu et al., 2017). HWE has been used as a method of removing wrongly genotyped SNPs from analyses (Wittke-Thompson et al., 2005, Fajardo et al., 2012, Pongpanich et al., 2010, Waples, 2015). It has also been used to identify regions associated with phenotypes, particularly those under strong selective pressure (Piel et al., 2016, Nielsen et al., 1998, Lachance, 2009, Alvarez, 2008). As this work is looking for alleles that may affect phenotypes relevant to production and the population in question is under strong selection for productive traits, using this as a filter may remove relevant loci. However, unsurprisingly, prioritising

based on deviations from HWE appears to primarily identify false-positives. Additionally, HWE assumes a large, randomly breeding population (Waples, 2015, Wittke-Thompson et al., 2005), but the population in question is not breeding randomly and consists of a small subset of individuals from multiple generations of a line under strong selective pressures. Therefore this method may not be appropriate in filtering or prioritising in this particular set of individuals. Subsequently this was not used to filter or prioritise. The variant in SERPINA3 was retained through all of the additional filters, was subsequently validated and remained of interest when phenotype data was taken into account. Thus in this particular case, the deviation from HWE may be meaningful and may relate to selective pressure. However this deviation may also relate to genetic drag and further work would need to be done to determine the cause.

Filtering based on gene function was necessary to reduce the list of variants down to a workable number. The removal of genes related to olfactory function greatly reduced the list of variants, and it is unlikely that any of these variants have an impact on the survival of the individual. Variant calls in olfactory genes are very common in pigs, as pigs naturally rely on their sense of smell to find food and they have one of the largest numbers of functional olfactory receptor genes of any known species (Nguyen et al., 2012), and these are generally not under selective constraint and tend to occur in poorly assembled clusters in the genome. Filtering by gene function is greatly limited by the

annotation of the genome, which is in turn limited by the contiguity of the assembly in genic regions. Filtering out variants that are not in named genes may remove potentially interesting variants. However it is beyond the scope of this work to validate each of the variants in these unnamed genes.

Despite the strict filtering that was carried out on these data, there were still false-positive calls remaining. Each variant was individually visualised to confirm the variant call, and 6-frame translations were carried out to confirm the consequence. While some of the prioritised genes were fragmented and not named in Ensembl, 6-frame translations of the sequence leading up to the variant aligned well to the respective proteins, and the consequence of the variants were supported by this. However, it should be noted that NPM1 is annotated elsewhere in Sscrofa10.2 and ENSSSCG00000004427 may be a paralog or pseudogene – however Ensembl also has supporting evidence for a protein aligning here. Following this *in-silico* validation, further validation was carried out to confirm the variants exist in the individuals through amplicon sequencing. In the majority of cases they were validated with only three of the 22 failing to validate. Two of the non-validated variants were later confirmed to exist in another set of pigs.

The majority of the genes prioritised from the filtering of the exome data have been implicated in cancer: MST1R (Catenacci et al., 2011, MOSER et al., 2012), RPLP1 (Perucho et al., 2014), NPM1

(Grossmann et al., 2011), CFTR (Than et al., 2016), CPT1A (Pucci et al., 2016), IL1RAP (Ågerstam et al., 2015), LAMA4 (Wragg et al., 2016), LAMB3 (Jung et al., 2018), LRRFIP2 (Thorsen et al., 2008), RDM1 (Li et al., 2017c), SERPINA3 (Yang et al., 2014), TGFB2 (Lebrun, 2012, Yeung et al., 2013), TRRAP (McMahon et al., 1998), and VCAN (Yeung et al., 2013). Often cancers involve aberrant expression of developmental genes. Tumorigenesis includes increased cell proliferation, tissue invasion and angiogenesis, much like the processes involved in embryo growth and implantation (Gailani and Bale, 1997, Fukuda et al., 2008). However, this may be a reflection of the volume of research carried out on cancer, and association with more specific reproductive phenotypes are probably a better indicator of relevance here.

From the 21 candidate genes identified from filtering the exome sequencing data, 11 have been directly associated with embryonic lethality in other species. These include CFTR (Lu et al., 2012), CPT1A (Nyman et al., 2005), MST1R (Muraoka et al., 1999), VCAN (Hatano et al., 2012), WDR36 (Gallenberger et al., 2011), NPM1 (Grisendi et al., 2005), TPT1 (Chen et al., 2007), MTHFD1 (MacFarlane et al., 2009), RPLP1(Perucho et al., 2014), TRRAP (Herceg et al., 2001) and TGFB2 (Memon et al., 2008). Most of these genes are associated with very early embryonic death, with the exception of NPM1 and TPT1, which have functions in determining embryonic stem cell fate and tend to cause death post-implantation in mice (Johansson and Simonsson,

2010). In cattle, during the maternal-to-embryo transition of the transcriptome, there is an almost 3-fold increase in TPT1 expression when the foetal genome is activated (Vigneault et al., 2009). RPLP1 knockout in addition to being embryonic lethal is also associated with several phenotypes in heterozygotes, including reduced body size, male infertility, systemic abnormalities and an increased frequency of post-natal death (Perucho et al., 2014), additionally there is evidence that this gene is expressed abundantly in the bovine foetus (Muramatsu et al., 2002). RPLP1 is known to be essential for normal development of the nervous system in embryos (Perucho et al., 2014) and the protein has an important role in the elongation step of protein synthesis and has been directly implicated in increased cell proliferation and the bypassing of replicative senescence of cells (Artero-Castro et al., 2009)

Of the genes associated with embryonic lethality, the two missense variants in CFTR are of particular interest, these were included in the final list as a number of variants in this gene are associated with cystic fibrosis in humans and do not need to knock out the gene to cause the phenotype (Bobadilla Joseph et al., 2002). Of over 1,900 CFTR mutations reported in humans, over 1,500 are considered potentially causative of cystic fibrosis symptoms, and 16.4% of Europeans carry at least one protein truncating allele in the gene (De Boeck et al., 2014). In mice, while disruption of CFTR is often lethal, some individuals do survive, with roughly 0.23 out of every 4 individuals from a heterozygous cross being homozygous for the variant. It has been

suggested that maternal CFTR mRNA stored in the oocyte may be sufficient for an embryo to survive. This may also explain how cystic fibrosis patients with two cystic fibrosis alleles are able to survive gestation (Lu et al., 2012). Human males with cystic fibrosis are usually infertile, due to a failure in the development of the vas deferens (Xu et al., 2007). In pigs, CFTR is highly expressed in the oviductal epithelium. Porcine CFTR mRNA is 93% identical to human CFTR mRNA, whereas murine CFTR mRNA is only 78% similar (Chen et al., 2010). A porcine model for cystic fibrosis has been developed by Welsh et al. (2009) and was found to have symptoms more similar to the human disease than previous mouse models. However the authors do not describe any reproductive phenotypes. The two missense variants identified here may not cause a phenotype as severe as embryonic lethality or absent vas deferens, but variants that modify the protein have the potential to reduce its functionality and missense variants in humans have been identified that cause cystic fibrosis ranging from a mild phenotype to severe cystic fibrosis (Krasnov et al., 2008). Pigs with a mild phenotype may be less productive than healthy pigs. Additionally it may be useful to have medical models that can mimic these different levels of severity.

Some of the other prioritised genes have previously been associated with other phenotypes linked to reproduction, and these are described in table 3-7.

Some of these may be particularly interesting, for example IL1RAP plays a role in maternal recognition of pregnancy and in pigs reaches

peak expression immediately prior to implantation during the elongation phase of the embryo. Its ligand, IL1, increases blood cell permeability and leukocyte extravasation (Mathew et al., 2016). In the animals sequenced here there were two individuals that had homozygous knockouts of IL1RAP present in the population, suggesting this is not a lethal variant.

**Table 3-7- Table describing genes containing short listed variants that have previously been associated with reproductive phenotypes**

| Gene | Associated phenotypes | References |
|---|---|---|
| IL1RAP | Maternal recognition of pregnancy | (Seo et al., 2011, Mathew et al., 2016) |
| LAMA4 | Implantation | (Shan et al., 2015, Malinda and Kleinman, 1996) |
| MST1R | Establishing maternal/foetal interface | (Muraoka et al., 1999) |
| NPM1 | Placental disease | (Grisendi et al., 2005) |
| SERPINA3 | Implantation, Placental disease, pre-eclampsia | (Chelbi et al., 2007, Chelbi et al., 2012, Altmäe et al., 2012) |
| MTHFD1 | Pregnancy loss, impairment of foetal growth, early separation of placenta | (Parle-McDermott et al., 2005, Parle-McDermott et al., 2005, Beaudin et al., 2012) |
| VCAN | Implantation | (Altmäe et al., 2012, San Martin et al., 2003) |

MTHFD1, while lethal in the embryo when knocked out, is also associated with a maternal phenotype in humans when missense polymorphisms are present, polymorphisms in this gene in the mother are associated with unexpected pregnancy loss in humans (Parle-McDermott et al., 2005), impairment of foetal growth in mice (Beaudin et al., 2012), and early separation of an otherwise healthy placenta from the uterus in humans (Parle-McDermott et al., 2005). Changes in SERPINA3 expression have been associated with a number of placental diseases in humans, with overexpression believed to decrease cell adhesion in trophoblasts. Maternal alleles of SERPINA3 have been associated with intra-uterine growth restriction and preeclampsia (Chelbi et al., 2012, Chelbi et al., 2007).

Several genes from the prioritised list also have phenotypes linked to traits not directly related to early embryo survival, but may play a role in defects and mortality later in gestation or life. Two of the prioritised genes identified, EXOG (Tann et al., 2011) and RDM1(Jaroudi and SenGupta, 2007), are involved in the repair of breaks in DNA with the latter being linked specifically to this process in the embryo. While neither of these genes have documented lethal knockout phenotypes, EXOG depletion is associated with cell death through rapid accumulation of single stranded DNA breaks in the mitochondria (Tann et al., 2011, Tigchelaar et al., 2014). However, the EXOG stop gain was found in the homozygous form in an adult female, so can be compatible with life.

While 54 sows were selected for further sequencing, only 27 of these had tissue samples available, limiting any further investigation into the role of the candidate variants in these underperforming litters. Additionally, most of these sows are from different lines than the boars, meaning any changes in allele frequency may be from differences in lines rather than a higher prevalence of a variant in dams of underperforming litters. The indel in ENSSSCG00000012880 (CPT1A) had the clearest increase in allele frequency, with only six of the 96 boars carrying the variant and 25 out of 27 of the sows carrying the variant. Of the three sows that were from the same line as the boars, two carried the CPT1A variant. Unfortunately, CPT1A is fragmented in Sscrofa10.2 with part of the gene (ENSSSCG00000012881) on the opposite strand. This makes it difficult to confirm the consequence of the variant and increases the chances of mismapping causing a false-positive call. CPT1A has a role in the metabolic function of the embryo (Nyman et al., 2005) and it would be interesting to see the impact of this variant on a more complete gene model to see if it is likely to cause disruptions to the protein function.

Four of the variants were found to be homozygous in some of the females suggesting these are not lethal, although the percentage dead in litters of the females homozygous for a stop-gain SNP in EXOG and a missense SNP in CFTR, respectively, were notably higher than average. While this may suggest a maternal phenotype, more data would be needed to test this. As has already been discussed, it has

been suggested that maternal CFTR is highly expressed in the porcine oviductal epithelium and may play an important role in early embryo survival. The litter from the dam homozygous for a CFTR missense variant had a small litter size of 9 with only 2 individuals surviving until birth. While it cannot be concluded that this variant is the cause of the poor performance of this litter, it may be worth following up on this and identifying more homozygous individuals in the population. It would be unusual for a seriously detrimental variant to be so prevalent, however it is possible that in the heterozygous form there is some kind of survival advantage. The prevalence of CFTR mutations in humans has been proposed to be as a result of heterozygous individuals being more resistant to cholera than the general population (Rodman and Zamudio, 1991).

Heterozygous crosses were found for four variants in EXOG, LRRFIP2, SERPINA3, and LAMB3, respectively. Ideally the litters of these crosses could be genotyped to identify any missing homozygosity, but unfortunately tissue samples are not available for all of the individuals from each of the litters. Of the data for the heterozygous cross litters available here, the variant in SERPINA3 was associated with the highest number of dead piglets at an average of 5 per litter which is over 4 times that of the full set of litter data available for the selected boars.

The follow up sequencing of the sows was very limited due to low availability of samples and time restrictions. Ideally this would be done

with an equally representative set of females across all of the variants, and would include more individuals including the dams of both the worst litters and the best litters. With so few heterozygous crosses identified and so few females from the same lines as the boars it is very difficult to draw any conclusions.

### 3.4.2 GWAS

A GWAS was used to find variants associated with reproductive phenotypes. GWAS generally involves very large populations with genome-wide genotype data and is more successful when a small number of loci have a large effect on the phenotype (Stranger et al., 2011). In this case variant genotype data was available exome-wide for just 96 individuals, however, the individuals in question are from a much larger breeding population where for many individuals SNP chip genotyping data and phenotype data are available. The 96 exomes were used as a reference population to impute variants into 351 additional individuals and GWAS was performed against two phenotypes: EBVs for number born to purebred (BTP; the number born in a cross between two individuals from the same line) and number stillborn (NSB). Owing to the fact the models used in calculating these EBV figures were designed internally by Genus PIC, further details cannot be discussed here. The phenotypes available for this work are only subtly different, however the lack of correlation between the two (figure 3-3) indicates they are distinct phenotypes. NSB likely indicates an increase in deaths late in gestation, or during farrowing, and may be

more likely to be caused by non-genetic factors on the part of the foetus such as physical trauma, disease, or resource competition through overcrowding of the uterus. A potential genetic cause from the foetus could relate to the formation of the placenta, although this can be lethal early in pregnancy. It can also cause conditions such as pre-eclampsia in humans and lead to restricted growth through inefficient interaction between maternal and foetal tissue. NSB may also be a phenotype of the dam, as struggling or delays during farrowing may kill the piglet at or near parturition, often through asphyxiation (Alonso-Spilsbury et al., 2004). BTP may indicate death earlier in gestation, when it is still possible for the dam to resorb the foetus. BTP may share some causes with NSB, but also includes foetuses that are not viable due to genetic defects, and zygotes with severe defects that may not survive until implantation. Additionally, BTP may be a phenotype of the dam with smaller litters being a consequence of the uterine environment or ovulation rate, for example.

The accuracy of the imputation in this project would likely be higher with a larger reference population to increase the number of represented haplotypes, however this could also be increased by targeted sequencing of focal individuals which together cover the majority of the haplotypes present (Gonen et al., 2017).The power for the GWAS analyses was very low owing to the small number of animals with genotype data made available for the project. However, commercial breeding companies typically have more than enough individuals

genotyped to drastically improve the power of an association of this kind. Unfortunately, of the small number of genotyped individuals provided very few had EBVs of a high enough reliability to use in the association analysis, this is not ideal, and often GWAS with small sample sizes are enriched for false-positive associations (Hong and Park, 2012). GWAS from SNP chips rarely identify the causative variant, but one in LD with the causative variant, so exome variants in close proximity to significant variants were identified in addition to looking for candidate genes near significant regions. Whereas the previously discussed method of reducing the number of exome variants required stringent filtering that forced the analysis to focus on rare null variants, this method allows some more common, moderate impact variants to be looked at as potential candidates.

It is likely that the imputation of variants is less reliable in regions that are poorly assembled, and while the regions identified as LQLC were excluded from the imputation, contig breaks will still interfere with the identification of haplotypes. Additionally, exome variants are not likely to be as evenly dispersed across the genome as a high density SNP array, and this may cause some regions to be imputed less accurately than others. As not all of the exome variants will have imputed reliably, the regions surrounding the variants with the highest associations were searched for moderate or high impact variants which may be responsible for the association. Relatively few variants were found in the regions identified, this may suggest that if there is a causative

variant in these regions they may be intergenic and play some kind of regulatory role. Here only variants in and close to exons are available, however judging by the reduced search space afforded by the exome variants under significant peaks, whole genome sequencing used in this way may help to identify these, it would however be limited by the accuracy of haplotype imputation in a fragmented assembly. A consequence of the poorly assembled genome can be seen in figures 3-5 and 3-6, where a single variant in a region appears to be significant on chromosome 15, when its true genomic location is on chromosome 3. Similar problems have been identified by other researchers using GWAS on the pig genome, for example researchers identified a SNP on chromosome 4 in complete LD with a region on chromosome 14 (van Son et al., 2017, Yang et al., 2013).

The small reference population and small sample size combined with the unreliable genome assembly increase the likelihood that the results of this GWAS are unreliable. However some of the genes identified may be viable candidates. For several of them the literature search identified functions that may be relevant to these phenotypes, though further work would be needed to validate these and all results should be treated with caution.

For BTP, no exome variants were found that looked like candidates for the phenotype, however several genes were identified. Most of these genes do not appear to be relevant to the phenotype and those with

**Table 3-8- Table summarizing the known functions and phenotype associations of genes located close to significant peaks in the GWAS**

| EBV | Gene | Reported function or phenotype associations | References |
|---|---|---|---|
| BTP | ELMO2 | Intraosseous vascular malformations | (Cetinkaya et al., 2016) |
| BTP | COL11A1 | Skeletal dysplasia, bone microarchitecture abnormalities | (Annunen et al., 1999, Griffith et al., 1998, Hafez et al., 2015) |
| BTP | CEP120 | Centrosomal protein associated with ciliary function | Shaheen et al. (2015) |
| BTP | SLC35C2 | Notch signalling | (Lu et al., 2010) |
| BTP | ZEB1 | Female fertility | (Hasuwa et al., 2013, Li et al., 2015, Jimenez et al., 2016) |
| NSB | ESR1 | Litter size | (Li et al., 2017b, Munoz et al., 2007, Rothschild et al., 1996, Short et al., 1997, Alfonso, 2005) |
| NSB | PARVB | Adhesion, angiogenesis, implantation | (McCallie et al., 2016) |
| NSB | GABRA2 and GABRA4 | Brain development and function | (Wu and Sun, 2015, Varshney et al., 2017) |
| NSB | ARHGAP24 | Angiogenesis, growth, close to QTL for prenatal survival in pigs | (Meng et al., 2017b, Su et al., 2004, Hernandez et al., 2014) |

| | | | |
|---|---|---|---|
| NSB | MYOM2 | Unknown, expressed in mouse embryo | (Ringwald et al., 2000) |
| NSB | ARHGEF10 | Developmental myelination of peripheral nerves | (Verhoeven et al., 2003) |
| NSB | SERPINB cluster | Various | (Heit et al., 2013) |
| NSB | SERPINA cluster | Various | (Heit et al., 2013) |

known functions are listed in table 3-8. These are likely to be enriched for false positives due to the low number of individuals.

ZEB1 has relevance to the phenotype as it has a role in several stages of female fertility including ovulation (in mice; Hasuwa et al., 2013); endometrial development and implantation (Li et al., 2015, Jimenez et al., 2016); and uterine quiescence and contractility (Renthal et al., 2010). While ZEB1 is the most likely candidate out of those associated with BTP  the region is only approaching significance in the exome imputed dataset, but is significant with SNP chip data alone. If this is not a false signal, it may suggest that the most significant association lies within a region outside of the exome that may regulate this gene.

For NSB, three variants were annotated as high impact, however two of these were in fragmented genes and it is unclear if the annotation reflects the true impact of these variants, and the third variant was homozygous in all individuals and likely represents an error in the reference or a benign variant fixed in the population. Three missense mutations that were not SIFT tolerated were identified, one of these was in a fragmented gene, and again the annotation may not be correct. The other two are in CORO2B and SERPINB8. CORO2B is from the Coronin family of actin regulators, but its role is not well defined, it is thought to be involved in reorganisation of the ventral neuronal actin cytoskeleton and cell migration (Chan et al., 2011, Rogg et al., 2017). There is no specific indication that CORO2B would interfere with reproductive traits or survival during gestation. SERPINB8 is part of the

SERPIN superfamily along with SERPINA3, a gene discussed previously. Both the SERPINA cluster and the SERPINB cluster appear to show some significance with NSB. The proteins in the SERPIN superfamily share a similar structure, but are functionally diverse, and many of the proteins are not well understood. Unfortunately, SERPINB8 is not well characterised (Heit et al., 2013), however it has been shown to have an inhibitory activity towards furin (Leblond et al., 2006), and has been linked to Exfoliative Ichthyosis (Pigors et al., 2016) or “Peeling Skin Syndrome” in humans. Again, there appears to be no reason to believe this gene should cause an increase in stillbirths with the information currently available. SERPINA3, as discussed previously, is a better candidate for reproductive phenotypes.

A number of genes appear to be associated with NSB, and these are summarized in table 3-9, many of them do not appear to be relevant to the phenotype, but some of them are supported by previous research . For example, a marker in an intron of porcine ESR1 was initially associated with litter size by Rothschild et al. (1996) who found an additive effect of 0.6 piglets/allele in Large White pigs from Genus PIC lines, and these pigs may have links to the pigs used in the work in this thesis. Since this initial discovery there have been a number of papers finding similar results (e.g. Short et al., 1997) and contrasting results (e.g. Dall’Olio et al., 2011), however a meta-analysis of 15 published studies including data on 9,329 sows supports an additive effect of the marker on litter size and number born alive (Alfonso, 2005). A number

of hypotheses have been presented for the contrasting results in the literature including interactions between genotype and environment, epistasis, or the gene not being directly causative of the phenotype but linked to it, with the different haplotypes in different populations disrupting the link. More recently a synonymous substitution in exon 5 of ESR1 has been identified as being associated with an additive effect on litter size in a Chinese-European pig line (Munoz et al., 2007), while unlikely to be causative, it is a variant from a different population in the same region with the same effect as the previously identified polymorphism. This suggests these variants are linked to a causal variation, though this has yet to be identified. The most significant variant in that region from this work is rs321737372, a missense variant in a gene (ENSSSCG00000004082) with no clear orthologs. A significant region on SSC8 overlaps a previously identified QTL associated with number stillborn by Schneider et al. (2015), this region contains GABA receptors. GABRA4 is a GABA-A receptor, as is nearby GABRA2, these genes are important in brain development and function (Wu and Sun, 2015, Varshney et al., 2017), knockouts of other GABA receptors in mice are viable (Rau et al., 2009, Chandra et al., 2005). Another significant region on chromosome 8 is in the gene ARHGAP24, which has previously been associated with growth in pigs (Meng et al., 2017b). Additionally, it is close to a region previously associated with prenatal survival in pigs (Hernandez et al., 2014). The significant region on SSC15 overlaps a previously identified QTL associated with number

born alive by Onteru et al. (2012), although the only candidate gene for the region mentioned was “novel protein”. In this case the significant region occurs over a fragmented MYOM2. MYOM2 is a muscular component and very little work has been done on this gene. MYOM2 is one of the genes which contained a SIFT deleterious missense mutation in the exome dataset, unfortunately the fragmented state of the gene in the assembly makes this call unreliable.

The SERPINA3 stop-gain SNP may be the most interesting variant identified, with the observed litters from crosses between heterozygous individuals having a very high percentage dead at 40% of the litter, on average 5 dead piglets per litter. The SNP has a similarly high minor allele frequency in both the sequenced boars and the sows, but has not been observed homozygous. Additionally, there is a peak at the SERPINA cluster in the GWAS for NSB. The variant is overlapped by a haplotype that has previously been associated with smaller litter sizes in sire x maternal grand-sire carrier crosses, and the observed genotypes were significantly different from the genotype distribution predicted by the Hardy-Weinberg Equilibrium for the exome sequencing data. The gene’s known associations with placental disease in humans and mice support this variant’s potential to impact reproductive phenotypes in pigs.

### 3.4.3 Conclusion

While filtering out the low confidence regions of the genome did improve the accuracy of variant calls from the exome sequencing data,

it is clear that false-positives are still abundant. Additionally, throwing out data for over a quarter of the genome is not ideal and likely removes good candidates along with low-quality calls. The consequences of a poorly assembled genome still impact on the data even after removing the previously identified low-quality regions. While GWAS is fairly robust to small errors in the genome, large inversions, missing sequence and wrongly ordered contigs break the haplotypes needed to accurately impute variants and obscure signal in association analyses.

Both of the methods of identifying candidate variants and genes described here found plausible candidates, however further work needs to be done to confirm the presence of the variants, their impact on the protein, and whether or not a phenotype can be observed. Ideally, a larger population would be sequenced to increase the chances of observing rare, but non-lethal variants in the homozygous state. This would allow for better filtering of the variants which may also allow for investigation of variants predicted to be of moderate impact in addition to the high impact variants. For the GWAS, a larger population would increase the power and accuracy of the analysis. Unfortunately the power of the GWAs presented here is very low, and while some of the regions identified may have some relevance to the phenotype and association cannot be confirmed. WGS rather than WES would allow for investigation of candidates in regulatory regions, however the cost of this would limit the number of individuals that could be sequenced and would likely result in a similarly small reference population for

imputation. An alternative approach being proposed by Ros-Freixedes et al. (2017) is to use low-coverage WGS of a large number of individuals as a strategy to assemble the haplotypes present in a population rather than only using one small set of individuals sequenced at depth as a reference group. This would increase the accuracy of imputation to the test population.

Assuming a larger population reduced the list of candidates further, crosses could be done between heterozygous individuals to identify any reduced litter sizes or unexpected genotype distributions in the resultant litters. To an extent, this could be done through the records of a pig breeding company by identifying individuals that carry the variant and have already been crossed. However, this depends on the tissue/DNA storage practices of the company. Such records could likely only reveal changes in litter sizes as availability of tissue samples to genotype from entire litters is unlikely. For some of these variants, particularly those in genes with functions related to implantation such as SERPINA3, looking at other phenotypes such as weight at birth may be of interest as it is known that in humans problems at the maternal/foetal interface can result in reduced growth (Pardi et al., 2002). Lighter birth weight pigs take longer to reach market size than average (Beaulieu et al., 2010).

For the variants in CFTR, one of the variants was found in the homozygous form in a number of individuals' exome data, and the other was only found heterozygous. If pigs homozygous for each of these variants are identified they can be phenotyped for known symptoms of

cystic fibrosis, or a deep phenotyping approach could be taken (Robinson, 2012). Alternatively, these homozygous pigs can be created through crossing heterozygous individuals, which would be easier to find in the population. Cystic fibrosis in human is most commonly found in individuals with a deletion of a single amino acid in exon 10 (Boyle and De Boeck, 2013), however there are hundreds of other causative variants causing disease of varying degrees of severity. It is possible that one of the variants identified here mimics some form of human disease. The less common of the two variants was only seen once in the homozygous form, and this individual's litter had a very low success rate with 9 individuals born, but only 2 alive. Many things could have caused this poor performance, however identification of more homozygous individuals and phenotyping them for reproductive traits would help to understand any role this variant may play.

In *Litopenaeus vannamei,* LRRFIP2 knockout is associated with increased mortality on exposure to certain pathogens (Zhang et al., 2013), so this variant may be of interest in studies of disease resistance or tolerance in pigs. The gene ENSSSCG00000004082 that is located close to ESR1 on SSC1 contains the most significant variant of the region, but the gene has no identified function and may not be reliable. Previous research has focussed on ESR1 as the major candidate of the region. As this region has evidence from multiple sources of an association with reproductive traits, it warrants further investigation. With the previous work that has been done and has failed to find a

causative variant, the phenotype may be caused by a structural variant which are harder to identify using the data that has already been applied to the problem, or a variant outside of the exons that regulates expression of ESR1. Identification of the causative variant may also have been hampered by the assembly of the ESR1 gene, which in Sscrofa10.2 is fragmented. The use of long-range PCR across the region accompanied by short-read and long-read sequencing of individuals with the causative haplotype may aid in the identification of the cause in this region. Short-reads are currently necessary to identify SNPs and indels that may have a regulatory effect, but long-reads would be more successful at identifying structural variants. The region containing ARHGAP24 has also previously been associated with reproductive phenotypes, and a targeted approach to investigate this gene and the surrounding region may be warranted. MYOM2, while not associated with phenotypes in previous studies, was in the most significant region in this work, why this variant should impact the phenotype is unclear and further work on understanding the gene may help to elucidate this.

Ultimately, in order to prove causation for any of these variants, knockouts of the gene could be created in cells or pig embryos using gene editing technologies such as CRISPR/Cas9 to directly observe the resultant phenotype. Following confirmation of a detrimental variant this can be included in the selection criteria for breeding programs. Due to

the large litter sizes and short generation intervals in pigs changes such as this can be implemented in the nucleus herd relatively quickly.

This chapter has focussed on identification of candidate variants that may be related to embryonic lethality or reduced reproductive performance. Using exome sequencing data, two approaches (identification of high impact, protein truncating variants and GWAS) were used to identify both rare and common variants that might be of importance to commercial pig production. The quality of the reference genome is a challenge both for variant calling methods and SNP-chip based GWAS, with the work here excluding over a quarter of the exome due to quality concerns. While a number of candidates have been reported here, further work is needed to confirm whether or not these are causative of any observable phenotype. While this work was limited to the exome, costs of WGS continue to decrease and in the future similar work including the whole genome may be feasible. Importantly, this would allow for regulatory regions to be explored further. The identification of variants related to important economic traits in the pig has the potential to increase productivity in the pork industry and with further work some of the candidates presented here may become valuable selection targets.

# CHAPTER 4: TRIO SEQUENCING TO SEARCH FOR CANDIDATE VARIANTS FOR FOETAL MUMMIFICATION IN PIGS

---

*"What do your parents know, about surviving? "*
- Lemony Snicket, The End

## 4.1 Introduction

While the stringent filtering strategy used in chapter 3 can address some of the limitations of current genomic resources, it does so at the expense of disregarding a large portion of the exome. Additionally, by sequencing only the exome, variants in regulatory regions are less likely to be identified. In this chapter, an alternative strategy aims to include more of the genome by instead using whole genome sequencing and filtering primarily based on the variants and genotypes that are present in the sire and dam of an affected individual to identify candidates for a major trait of interest: embryonic lethality.

In production animals, reproductive traits are an important target for selection. Females are bred to produce as many healthy offspring as possible, increased litter size and reduced foetal mortality are therefore major targets. Foetal mortality may result in a number of outcomes, in monotocous species this will often result in resorption of the foetus or early evacuation of the uterus. However, in certain circumstances, particularly in polytocous species where other foetuses are still alive, it may lead to mummification. Foetuses that die after calcification undergo autolysis shortly after death, initially swelling as serous fluid accumulates before this fluid is resorbed by the uterus, leaving a dehydrated, shrivelled, but otherwise preserved foetus. Some causes of foetal mummification have been identified, these include viral infection (Love et al., 2008, Joo et al., 1976, Mengeling et al., 1975), parasitic infection (Dubey, 1999), ectopic pregnancy (Tena-Betancourt et al.,

2014), uterine torsion (Moore and Richardson, 1995, López and Carmona, 2010), and genetic or chromosomal abnormalities (Ghanem et al., 2006). The majority of cases of foetal mummification in pigs are attributed to management practices and resource limitations in utero (Wu et al., 1988, Lefebvre, 2015, Cozler et al., 2002). Viral infections are another major cause in pigs, including infections with Porcine Reproductive and Respiratory Syndrome virus (PRRSV -Terpstra et al., 1991, Plana et al., 1992) and Porcine Parvovirus (PPV -Joo et al., 1976, Mengeling et al., 1975). Limited research has been carried out on foetal mummification as it is rare in most domestic species, with the highest prevalence in swine (Lefebvre, 2015, Christianson, 1992) with estimates ranging from 1.5% to 6.8% of foetuses (Cozler et al., 2002, Schneider et al., 2012, van der Lende and van Rens, 2003, Wu et al., 1988). Schneider et al. (2012) estimated a low heritability in a genome-wide association study for foetal mummification in pigs at 0.06 ± 0.04. However, estimates of prevalence and heritability are challenging, as smaller mummies may be lost during farrowing leading to an inaccurate count. Additionally, unidentified disease in a herd and varying management practices between farms may cause variability in prevalence. The definition of “mummified” and “stillborn” may also vary between farms.

The different estimations of prevalence between different studies may suggest there is a different prevalence between different breeds, although this may relate to the varying litter sizes and uterine space between different breeds (Wu et al., 1988). A study on mummified foetuses in cattle targeted a specific gene with a SNP known to cause foetal mortality in that species, they

found two out of ten mummified foetuses were homozygous for the causative SNP and an additional mummified foetus appeared to be missing the X chromosome (Ghanem et al., 2006). These results from mummified cattle suggest that there may be some genetic causes of mummification, though not necessarily a single cause for all cases. This combined with the wide range of non-genetic factors involved in foetal mummification may obscure any causative loci in a genome-wide association study. A large study by Derks et al. (2017) looked at 24,000 individuals from three populations genotyped for 80K SNPs and identified missing or depleted homozygous haplotypes across the populations, they found a number of regions of the genome that were associated with reproductive traits including 5 regions associated with increases in number of mummified foetuses. One of the regions identified by Derks et al. (2017) increased the number of mummified foetuses per litter 5-fold in carrier cross matings and the authors proposed BMPER as a candidate gene for this region. They later discovered a large deletion in a neighbouring gene BBS9. The deletion in BBS9 had a positive effect on the growth of the pigs, however it also deleted a promotor for BMPER, which is lethal when homozygous (Derks et al., 2018). The study benefited from having a very large sample size to identify regions and demonstrated a clear genetic cause for mummification that persists in a population of pigs owing to balancing selection. Mummification may be caused by a single variation as was found in that study, or it may be a number of loci together affecting either the foetuses survivability or the dam’s reproductive fitness.

In human research, it is recognised that miscarriage does not lend itself well to association studies, the results of which are poorly replicated. These are likely affected by low sample sizes, differences in study design and patient ethnicity, and poorly defined phenotypes. This is further complicated by differing and multifactorial aetiologies based on the stage of pregnancy the loss occurred, patient history of recurring miscarriages, and environmental factors (Rull et al., 2012). While the majority of sporadic pregnancy losses are attributed to foetal chromosomal abnormalities (Philipp et al., 2003, Menasha et al., 2005), these are linked to a relatively small proportion of recurrent miscarriages. Research has shown that in couples that are first degree blood relatives, miscarriages are two to seven times more common than the general population (Christiansen, 1996), and the siblings of recurrent miscarriage patients are twice as likely as the general population to suffer a miscarriage (Kolte et al., 2011), suggesting a heritable genetic component. Variants and haplotypes have been successfully linked with increased risk of miscarriage in humans (e.g. Ober et al., 2003, Almawi et al., 2013, Misra et al., 2016).

Exome sequencing has proven to be a valuable tool for affordable sequencing of large cohorts to identify candidate variants for a range of phenotypes where WGS is prohibitively expensive (Warr et al., 2015a). However both WES and WGS can be used in a more targeted way for identifying the causal variation underlying specific traits or phenotypes. In trio sequencing, the affected individual and two unaffected parents are sequenced, occasionally also sequencing one or more affected or unaffected

siblings. There are a number of commonly described inheritance models including dominant, recessive, additive, codominant and X-linked (Laird and Lange, 2011). In a dominant inheritance model, only one copy of an allele needs be present to affect a phenotype. Assuming both parents are unaffected this could only occur as a *de novo* variant in the affected offspring, detrimental dominant variants are unlikely to persist in a population under strong selection for productivity. A recessive variant will only affect a phenotype if two copies of the allele are present in an individual. In this case, both unaffected parents are heterozygous and roughly a quarter of each litter would exhibit the phenotype. An additive variant would increase the chances or severity of a phenotype with each additional allele depending on penetrance. In this case the phenotype would be most likely with the allele in the homozygous state, but may also be seen to a lesser degree in heterozygous individuals. In a codominant model both alleles are fully expressed and both influence the phenotype. X-linked variants may be dominant or recessive, although if both parents are unaffected, the variant cannot be dominant as at least one parent would be affected. A recessive variant on the X-chromosome of a healthy dam passed down to a male foetus will cause a phenotype owing to males only having a single copy of the chromosome. Trio sequencing assumes a monogenic, recessive inheritance model, but may also identify variants with additive or X-linked inheritance.

As only a small number of individuals need to be sequenced in trio sequencing, the costs associated with whole genome sequencing (WGS) are

more manageable than research involving large cohorts. While WES and WGS can produce huge lists of variants to annotate and filter, trio sequencing can substantially reduce the number of candidates, as those that are present in the same genotype in a known unaffected individual can be excluded as causative. Additionally, in cases where the parents are unaffected and assuming a monogenic, recessive inheritance model, the causative variant will be carried in the heterozygous state by the parents and the homozygous state by the offspring, or will be a novel variant in the offspring. Often this leads to a shorter list of variants that can be quickly filtered based on other criteria. The ability to filter variants based on the genotypes seen in the sire and dam of an individual offer an opportunity to identify candidate variants without the extreme filtering applied in other sequencing methods as is often necessary in species with imperfect reference genome assemblies and with little data available to assist in filtering. Trio sequencing is a common tool in paediatric diagnosis of rare diseases and has been used to diagnose a range of diseases (Lee et al., 2014, Meng et al., 2017a, Smith et al., 2014, Wang et al., 2011a). The practice is still relatively rare in animal science, though there are some examples using trio SNP genotyping (Lee et al., 2012), WES (Ahonen et al., 2013) and WGS (Sayyab et al., 2016, Reber et al., 2015). Often mixed sperm is used to inseminate females in an agricultural setting, and while it may not be possible to sequence a trio with an unknown sire, the sequencing of the dam and one or more healthy siblings can be used to assist in filtering of candidates to an extent.

A barrier to DNA sequencing for mummified foetuses is the quality of DNA that can be extracted from them. Mummification occurs in any conditions that halt autolysis, which includes the release of digestive enzymes from within the organs of the body and bacterial putrification. The process is stopped when the tissue water content drops below a critical level and inhibits bacterial putrification (Janaway et al., 2009). In the uterus of a healthy individual there are no bacteria to decompose the tissue, nor is there oxygen, and the foetus is naturally actively dehydrated after death by the dam. Prior to the dehydration of the foetus, autolysis from enzymes has already begun to damage the DNA, and the subsequent dehydration of the tissue may cause fragmentation of the remaining DNA (Zahradka et al., 2006). There are very few studies involving DNA extraction from mummified animal foetuses with most being carried out on human remains. In a study on mummified foetuses in cattle (Ghanem et al., 2006, Ghanem et al., 2005) researchers were able to obtain 1.1-3.2 µg/ml and although the authors do not state the elution volume whole genome amplification was required before PCR amplification of the target regions could be carried out suggesting a low total yield from this DNA extraction.. Conversely, in a case where the death of two young boys was molecularly diagnosed 40 years post mortem by extracting DNA from naturally mummified umbilical cords, researchers were able to extract 171.1 ng/µl and 59.1 ng/µl from the cords and diagnosed ornithine carbamoyltransferase deficiency via PCR amplification of the OTC gene (Takenouchi et al., 2014). Again, the authors do not state elution volume, but the kit used recommends 100-200 µl, this likely represents a higher yield

from these samples than the Ghanem et al. study. Umbilical cords benefit from being virtually sterile and void of digestive enzymes, limiting the potential for autolysis. Assuming DNA can be obtained from the degraded sample, the damage to the DNA may cause sequencing errors and variable or low coverage which can affect downstream variant calling (Parks and Lambert, 2015, Chen et al., 2017).

For this chapter DNA extraction and WGS was carried out on a trio of individuals: a mummified foetus, its sire and its dam. The mummified foetus was selected from a litter where a quarter of the foetuses were mummified during pregnancy, the expected number lost in Mendelian inheritance of a lethal recessive variant. Bioinformatic techniques were used to assess levels of DNA damage in the mummified foetus and to attempt to identify causative variants from the data. This method may be a valuable tool for identifying causative variants of specific phenotypes in pigs, just as it has been very successfully applied in paediatric diagnosis.

It should be noted that a larger sample size and selection of individuals from a specific pedigree suspected of carrying a variant associated with mummification, or investigation of a more well-defined monogenic trait would be more likely to be successful in identifying a candidate. However, this work serves as a proof of principal that the trio sequencing method can be used to greatly reduce a large pool of variants from whole genome sequencing to a manageable list of candidates for further filtering even when using a low quality reference genome.

The majority of the work was carried out by the author of the thesis, DNA extraction on the sires, dams and healthy siblings was carried out by Heather Finlayson at The Roslin Institute. Sample collection was carried out by the author, Heather Finlayson, Claire Stenhouse, and the staff at Dryden farm.

## 4.2 Methods

### 4.2.1 Sample collection

Samples were obtained from another project that was primarily concerned with foetal growth retardation (Stenhouse, 2017) in which mummified foetuses were to be discarded. Eight sows were inseminated using semen from two boars, with four sows per boar. Two sows were culled at day 60 of pregnancy for each boar, with the remaining four sows culled at day 90 of pregnancy. The reproductive tracts of the sows were removed intact, and the foetuses were removed from the uterus and whole mummified foetuses or samples of suspected resorbed tissue were collected. Additionally, one ear or leg from each apparently healthy foetus, blood from the sow and a semen sample from each boar was taken for DNA extraction. Foetus samples were labelled using the ID of the sow and the position of the foetus in the uterus, for example foetus 23976-L7 is from the left horn of the uterus of sow 23976 and was the seventh foetus in that horn. All samples were transported from the farm on dry ice and stored at -80°C prior to DNA extraction.

### 4.2.2 DNA extraction, clean up and sequencing

Unless stated otherwise, all tubes used were DNA LoBind tubes (Eppendorf; Hamburg, Germany), centrifuging was done at room temperature (RT) in a

Centrifuge 5415D (Eppendorf), or at colder temperatures as specified in a Centrifuge 5810R (Eppendorf).

### *4.2.2.1 DNA extraction from blood*

DNA was extracted from the blood of the eight sows as follows.

Stocks of two buffers were used which contained the following:

Buffer A (Red blood cell lysis buffer):
- 0.32 M sucrose
- 10 mM Tris HCl
- 5 mM $MgCl_2$
- 0.75% Trinton-X-100

Buffer B (Proteinase K buffer):
- 20 mM Tris HCl
- 4 mM $Na_2EDTA$
- 100 mM NaCl

8 ml of Buffer A was added to 8 ml of blood and 16 ml of cold Milli-Q (MerckMillipore; Massachusetts, USA) water in a 50 ml conical tube. The tube was inverted 8 times and incubated on ice for 3 minutes. The tube was centrifuged at 1137 x g for 15 minutes at 4°C. The supernatant was discarded and the pellet was resuspended in 2 ml Buffer A and 6 ml water by vortexing for 30 seconds. 5 ml of Buffer B and 500 µl 10% SDS was added to the pellet and it was vortexed for 30 seconds. 50 µl of Proteinase K (20 mg/ml) was added and the samples were incubated overnight at 55°C in a shaking water bath. The samples were allowed to equilibrate to room temperature and 4 ml of 5.3 M NaCl was added. They were then gently vortexed for 15 seconds. The tubes were centrifuged at 1485 x g for 20 minutes at 4°C and the supernatant was removed and retained. 13 µl cold isopropanol was added and the tubes were gently inverted 6 times. The DNA

was gently picked up using a wide bore tip and transferred to a 15 ml tube containing 5 ml 70% ethanol which was inverted 3 times. A wide bore tip was used again to pick up the DNA and transfer it to a 1.5 ml tube containing 1 ml 70% ethanol which was inverted 2 times. The samples were spun down at 10,000 x g for 5 minutes at room temperature and the ethanol was removed. The pellets were allowed to dry at room temperature for 15 minutes before being resuspended in 300 µl 10 mM Tris-HCl pH 8.0. Samples were left to resuspend for several hours before sample cleanup and QC.

### *4.2.2.2 DNA extraction from tissue*

DNA was extracted from 117 healthy foetuses as follows.

Tail buffer was prepared by combining 2.5 ml 1 M Tris HCl pH8.0, 10 ml 0.5 M di-sodium EDTA, 1 ml 5 M NaCl, 5 ml 10% SDS and 31.5 ml Milli-Q water. For each sample tail buffer plus PK was made by combining 600 µl tail buffer with 35 µl 20 mg/ml Proteinase K.

Tissue samples were cut on a chilled sterile petri dish using a scalpel into roughly 50mg pieces and then diced into smaller pieces and were immediately submerged in 635 µl tail buffer plus PK. The samples were incubated in a hybridisation oven at 55°C with rotation overnight. Samples were allowed to cool and 170 µl 5 M NaCl was added. The tubes were shaken and centrifuged at 16100 x g for 10 minutes. 500 µl of supernatant was removed from each tube and added to 1 ml of cold EtOH in a new tube, this was mixed by inversion. The samples were incubated at -20°C for 30 minutes and then centrifuged at 12000 x g at 4°C for 20 minutes. The

supernatant was discarded and the pellets were washed with 1 ml 70% EtOH and left for 5 minutes at room temperature. The tubes were centrifuged at 7500 x g for 5 minutes, the EtOH removed, and this wash was repeated. The sample was allowed to air dry for around 15 minutes and was resuspended in 500 µl 10 mM Tris-HCl pH 8.0.

### *4.2.2.3 DNA extraction from sperm*

Due to the high DNA compaction by protamines in sperm, standard DNA extraction methods cannot be used. A method based on that by Griffin (2013) was used as follows.

100 ml of sperm wash buffer was made by combining 1.5 ml 5 M NaCl, 1 ml 0.5 M EDTA pH8.0 and 47.5 ml Milli-Q water. 12 ml of extraction buffer was made by combining 8.48 ml 6 M guanidine thiocyanate, 240 µl 5 M NaCl, 400 µl 30% sarkosyl, 1.8 ml 1M DTT, 120 µl Proteinase K and 960 µl nuclease-free water.

The frozen sperm pellets containing roughly $10^9$ sperm cells each were resuspended in 0.5 ml sperm wash buffer and were then made up to 10 ml by adding more wash buffer. For each sample, 5 x 50 ml tubes of 2 ml of sample were used and were centrifuged at 750 x g for 10 minutes. The supernatant was decanted and the sample was vortexed to resuspend the cells in the residual buffer. 10 ml of sperm wash buffer was added and the sample was vortexed and centrifuged at 750 x g for 10 minutes. The supernatant was decanted and the sample was vortexed to resuspend the cells in the residual buffer. For each sample 1 tube was taken forward and

the other four sperm pellets were frozen at -80°C for future use. 6 ml of extraction buffer was added to each sample and the tubes were gently inverted several times to mix. Samples were incubated at 56°C on a shaking platform in a hybridisation oven for 2 hours, inverting the tubes 3 times half way through the incubation period. The tubes were allowed to equilibrate to room temperature. 4.8 ml isopropanol was added and mixed by gentle inversion until DNA strands could be seen. A wide bore pipette tip was used to transfer the DNA to a 2 ml tube containing 2 ml 0.1 M Na citrate in 10% ethanol and was incubated at room temperature for 30 minutes with occasional inversions and this was repeated by transferring the pellet to a fresh 2 ml tube containing 2 ml 0.1 M Na citrate in 10% ethanol. The pellet was transferred to a tube containing 1 ml 70% ethanol and the tube was inverted, this was repeated with a fresh tube of 1 ml 70% ethanol. The tube was then centrifuged at 5000 x g for 3 minutes at room temperature. The ethanol was removed and discarded and the samples were allowed to air dry for around 10 minutes. The pellets were resuspended in 800 µl 10 mM Tris-HCl pH 8.0. Samples were left to resuspend overnight before sample cleanup and QC.

#### *4.2.2.4 DNA extraction from mummified tissue*

A number of methods were attempted to extract DNA from the degraded tissue of the day 60 and day 90 mummified foetuses including different methods of sampling the tissue, different methods of lysing the tissue and different kits for extraction. The following method was the fourth iteration and found to be the most effective on the day 60 foetuses, which generally had

DNA that was more degraded than the day 90 foetuses. Day 60 foetuses were preferable as earlier deaths were deemed more likely to relate to causes other than restricted resources. Foetus 23982_R6 was chosen as it was a day 60 foetus and a quarter of the litter were mummified, in keeping with the expected distribution from a Mendelian inheritance pattern.

The Qiagen (Hilden, Germany) QIAmp DNA Investigator Kit was used for this extraction. The kit is designed for use in forensics and works well on degraded samples and small amounts of tissue. The workspace and two polystyrene boxes were cleaned thoroughly with trigene disinfectant and 70% ethanol. Dry ice was added to the two polystyrene boxes, one fully and one half full. To the half-filled box, sterile petri dishes and sterile scalpels were placed to cool along with the foetus which had been retrieved from -80°C freezer storage. The foetus was weighed on a scale and immediately returned to the dry ice, it weighed 6.175 g. The full box of dry ice was covered with aluminium foil and used as a work surface to keep the tissue cold. A chilled petri dish was placed on top of the foil and the foetus was placed in the petri dish. Two chilled sterile scalpels were used, one to hold the foetus still and one to cut it in half and scrape tissue from the inside of the cut. This method was used to minimise the potential contamination from the dam’s blood on the surface of the foetus. It had been impractical to clean the foetuses before storing them during sample collection. The target weight for the tissue samples was 10 mg per sample, however to reduce the risk of thawing due to over handling, larger weights were used and the lysate was split accordingly later. Three tissue fragments were placed in three 1.5 ml

tubes, these weighed 18 mg, 24 mg and 30 mg. 360 µl of ATL buffer from Qiagen QIAmp DNA Investigator Kit was added to the tubes with 18 mg and 24 mg in, and 540 µl was added to the tube with 30 mg in and they were allowed to equilibrate to room temperature before 40 µl Proteinase K was added to the tubes containing 18 mg and 24 mg and 60 µl was added to the tube containing 30 mg. The tubes were incubated in a hybridisation oven at 56ºC with rotation overnight. The lysate was split between 7 tubes by moving 200 µl from the first and second tube into fresh 1.5 ml Eppendorf tubes, and 2 x 200 µl from the third tube was moved to two new Eppendorf tubes so all 7 tubes contained 200 µl of lysate. The protocol for QIAmp DNA Investigator Kit "Isolation of Total DNA from Tissues" was followed from step 5 as follows. 200 µl of Buffer AL was added to each tube and mixed by vortexing. 200 µl 100% ethanol was added to each tube and immediately vortexed to homogenise. The tubes were left to stand for 5 minutes. For each sample, this mixture was pipetted into a QIAmp MiniElute column in a 2 ml collection tube and centrifuged at 6000 x g for 1 minute. The flow-through and collection tube were discarded and the collection tube was replaced with a new one. 500 µl Buffer AW1 was added to each tube and centrifuged at 6000 x g for 1 minute. The flow-through and collection tube were discarded and the collection tube was replaced with a new one. 700 µl Buffer AW2 was added to each tube and centrifuged at 6000 x g for 1 minute. The flow-through and collection tube were discarded and the collection tube was replaced with a new one. 700 µl of EtOH was added to each tube and centrifuged at 6000 x g for 1 minute. The flow-through and collection tube were discarded and the

collection tube was replaced with a new one. The tubes were centrifuged at 20000 x g for 3 minutes and the columns were transferred to 1.5 ml lo-bind tubes and incubated at room temperature for 10 minutes. 100 µl ATE buffer was pipetted directly onto the spin column membrane and incubated for 5 minutes at room temperature. The tubes were spun at 10000 x g for 1 minute. Any visible liquid remaining in the spin column was picked up using a P10 pipette and placed directly onto the membrane and the tubes were spun again at 10000 x g for 1 minute.

### *4.2.2.5 Quantification, quality control and RNase treatments*

The seven tubes of mummified foetus DNA were quantified using the Qubit (Life Technologies; California, USA) broad range (BR) kit. 1791 µl of BR buffer was combined with 9 µl BR reagent and vortexed thoroughly. 198 µl of this mix was added to 7 Qubit tubes for the samples and 190 µl of the mix was added to 2 Qubit tubes labelled standard 1 and standard 2. Samples were vortexed to homogenise and 2 µl of sample was added to each Qubit sample tube and 10 µl of broad range standard was added to the appropriate standard tube. The Qubit fluorometer was used to quantify the DNA. The quality was assessed using the Agilent (California, USA) Tapestation. A fresh Genomic DNA ScreenTape was removed from the fridge along with the Genomic DNA sample buffer and the Genomic DNA ladder and these were left at room temperature for 30 minutes. In a strip tube, 10 µl of Genomic DNA sample buffer was added to all of the tubes. 1 µl Genomic DNA ladder was added to the first tube and 1 µl of each sample was added to each subsequent tube. This was briefly spun down in a centrifuge before being

placed on the IKA (Staufen, Germany) vortexer for 1 minute at 371 x g. The tubes were again spun down briefly and were placed in the Tapestation and the Tapestation software was run.

The seven tubes of mummified foetus DNA were pooled into two tubes, one containing four samples and one containing three, 1 µl of Riboshredder (Epicentre; Wisconsin, USA) was added to each. The tubes were incubated at 37°C for 15 minutes. 800 µl of DNA binding buffer from the Zymo (California, USA) DNA Clean & Concentrator kit was added to the first tube and 600 µl to the second, to get a 2:1 ratio of DNA binding buffer to sample. The whole volume from both tubes was transferred to a new Zymo spin column in a 2 ml collection tube 800 µl at a time and it was centrifuged for 30 seconds at 12,000 x g and this was repeated until the full volume had passed through the tube. The flow through was discarded and 200 µl DNA Wash Buffer was applied to the membrane in the column. The tube was centrifuged for 30 seconds at 12,000 x g and the flow through was discarded, this DNA wash buffer step was repeated once. 200 µl of DNA elution buffer was applied to the membrane and it was incubated for 1 minute at room temperature before being transferred to a new 1.5 ml Eppendorf and being centrifuged for 30 seconds at 12,000 x g. Quantification with the Qubit BR kit was carried out as described above, scaled down to two samples worth of working solution. The quantification was done in duplicate and the Tapestation analysis was also repeated as described above.

Further steps were carried out for the sire and dam of the mummified foetus 23982_R6 only, “CB” and “23982” respectively, the other sires and dams and

the healthy siblings were not whole genome sequenced and these steps were not necessary. The DNA from CB and 23982 was roughly quantified using the Nanodrop (Thermo Fisher; Massachusetts, USA) and was diluted with 10 mM Tris-HCl pH 8.0 to 130 µg in 400µl. 1 µl Riboshredder was added to each and they were incubated at 37°C for 15 minutes. 2 Phase Lock Gel tubes (5PRIME; California, USA) were centrifuged at full speed for 1 minute to spin the gel to the bottom of the tubes. In a fume hood, 400 µl Phenol:Cholroform:Isoamyl was added to each Phase Lock Gel tube. Each sample was transferred into one Phase Lock Gel tube and the tubes were mixed by inversion until the contents were white and uniform, then they were gently mixed on a rotator for 5 minutes. The tubes were spun at max speed for 4 minutes at RT and the aqueous layer above the gel layer was transferred to a 2 ml tube. 16 µl of 5 M NaCl was added to each sample the samples were mixed with a Genomic DNA pipette tip. 920 µl cold ethanol was added to each sample and they were inverted until the DNA was visible. The DNA was picked up with a pipette tip and transferred to a new tube containing 1 ml 70% EtOH. The tubes were centrifuged at 16000 x g for 5 minutes at 4°C. The supernatant was removed and discarded and the pellet was left to dry at room temperature for 10 minutes. The pellets were resuspended in 300 µl 10 mM Tris-HCl pH 8.0. The RNase treated and cleaned DNA was quantified using Qubit as described previously, and the quality was checked using Tapestation as described previously.

#### *4.2.2.6 Sequencing*

Despite the mummified foetus sample having very poor quality DNA, it was sent to Edinburgh Genomics for sequencing using the TruSeq Nano (Illumina; California, USA) library preparation method on the Illumina HiSeqX. The resulting sequences were paired end with an expected mean insert size of 450 bp. Following initial QC of these data as described below, CB and 23982 were subsequently sequenced in the same manner.

### 4.2.3 Bioinformatic analysis

Unless stated otherwise, tools were run using default parameters

#### *4.2.3.1 Quality control and alignment of sequencing data*

FastQC (v. 0.11.5; Andrews, 2010) was used to check the fastq files for adapter content and basic sequence quality. The reads were trimmed using sickle (v. 1.33; Joshi and Fass, 2011) and FastQC was used again to confirm the trimming of low quality base calls. The fastq files were mapped to Sscrofa10.2 using BWA mem (v.0.7.15Li, 2013) and converted to sorted bam files with Samtools view and sort (v. 1.2; Li et al., 2009a). Additionally, Samtools flagstat was used to summarise basic mapping statistics. As the 23982_R6 was degraded, additional QC steps were taken. BEDtools genomecov (v. 2.26.0; Quinlan, 2014) was used to assess genome-wide coverage and check for a normal distribution, and coverage on the X and Y chromosome relative to the rest of the genome allowed for identification of the gender of the foetus.

### *4.2.3.2 Variant calling and filtering*

Picard MarkDuplicates (v. 2.7.1; http://broadinstitute.github.io/picard/) was used to mark duplicates in the three bam files. GATK HaplotypeCaller (v. 3.7-0; McKenna et al., 2010) was used to call variants, this was done using parameters --emitRefConfidence GVCF -variant_index_type LINEAR -variant_index_parameter 128000. Realignment and base recalibration were not done as GATK no longer recommends indel realignment with haplotype based callers, and it was decided that given the results of chapter 2, there is no reliable "truth set" of variants. Ensembl VEP (v. 89; McLaren et al., 2016) was used to annotate the variants. The variants were filtered down to those that were heterozygous in CB and 23982 and homozygous in 23982_R6 and were also annotated as high impact. Additionally, variants that were *de novo* in 23982_R6 were identified, those which were homozygous reference in CB and 23982 and homozygous alternative in 23982_R6. These *de novo* variants were visualised with Integrative Genomics Viewer (IGV; Thorvaldsdóttir et al., 2013) however, these are all likely to be false positives. This produced a short list of genes containing likely deleterious variants to be further investigated through a search of the literature.

BEDtools intersect (Quinlan, 2014) was used to check how many of the prioritised variants occur in low quality or low coverage regions identified in chapter 2. These regions were not filtered out in this case, but the annotated consequence for these variants are more likely to be unreliable.

In addition to the above, regions identified by Derks et al. (2017) as being associated with mummification were searched for variants that were heterozygous in 23982 and CB and homozygous in 23982_R6, that are annotated as having a moderate or high impact on the protein. Additionally, variants in the BMPER gene which was identified as a candidate in the study by Derks et al. (2017), was searched in the unfiltered, annotated VCF file for any variants that may impact the protein at any impact level and with any genotype.

### 4.2.4 Follow-up genotyping

Following identification of candidate variants during bioinformatic analysis, genotyping of the remaining sire and the seven remaining dams and 117 healthy foetuses was undertaken for four variants. All DNA samples were roughly quantified using the Nanodrop and samples were sent for genotyping with LGC (Middlesex, UK) group's genotyping services. All samples were over 100 ng/µl and 10 µl of each was sent along with the reference sequence of 400 bases upstream and downstream of the SNP from Sscrofa10.2 to LGC.

## 4.3 Results

### 4.3.1 Sample collection

Tissue samples were collected from 117 apparently healthy foetuses, along with blood samples from the eight dams and semen samples from the two sires. 17 whole mummified foetuses were collected with 10 from the day 60 litters and 7 from the day 90 litters. Each sire had a similar number of mummified foetuses with 8 from one sire and 9 from the other. The dams had between one and four mummified foetuses each, with the majority having only 1 mummified foetus recovered. The size of the foetuses varied with day 60 foetuses ranging from 0.5 cm-2.5 cm and day 90 foetuses ranging from 2.5 cm-10 cm. Examples of two day 60 and two day 90 mummified foetuses are shown in figure 4-1.

### 4.3.2 DNA extraction

Seven separate DNA extractions were carried out for the mummified foetus, 23982_R6, according to Qubit quantification the concentrations (ng/μl) of these were 27.8, 39.6, 43.4, 27.4, 29.8, 45.4, and 39.6 in 100ul each. The gel image from the Tapestation for the seven samples are shown in lanes B1-H1 in figure 4-2 in addition to an example of the electropherogram for one sample. Following pooling and RNase treatment, the sample was quantified with Qubit in duplicate, the concentrations were poor at 11 ng/μl and 13 ng/μl in 200ul total volume. The Tapestation was run on the pooled sample in duplicate and the results are shown in figure 4-3.

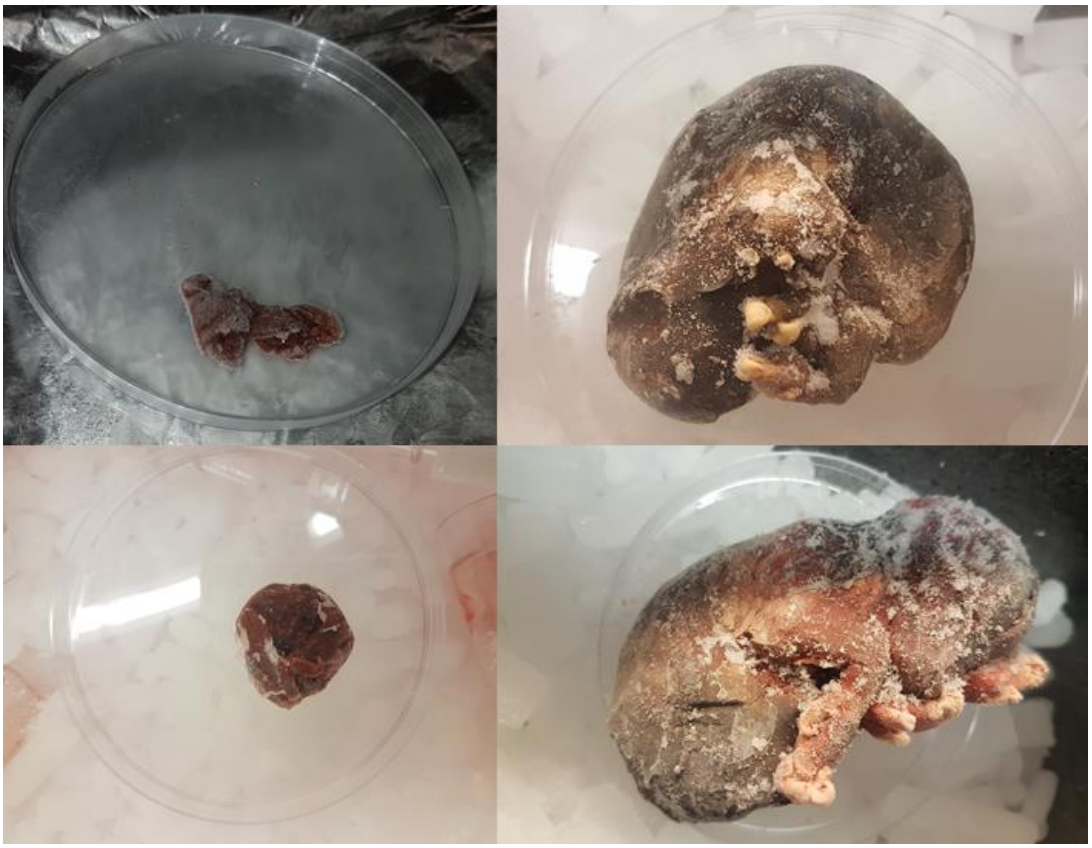


**Figure 4-1- Photographs of two day 60 mummified foetuses (left top, left bottom) and two day 90 mummified foetuses (right top, right bottom). All are in standard size petri dishes for scale.**

Extraction of high molecular weight DNA for the sires, dams and healthy siblings was successful, with all samples yielding >100 ng/µl in 100-200ul according to Nanodrop. The results of DNA extractions from CB and 23982 are described

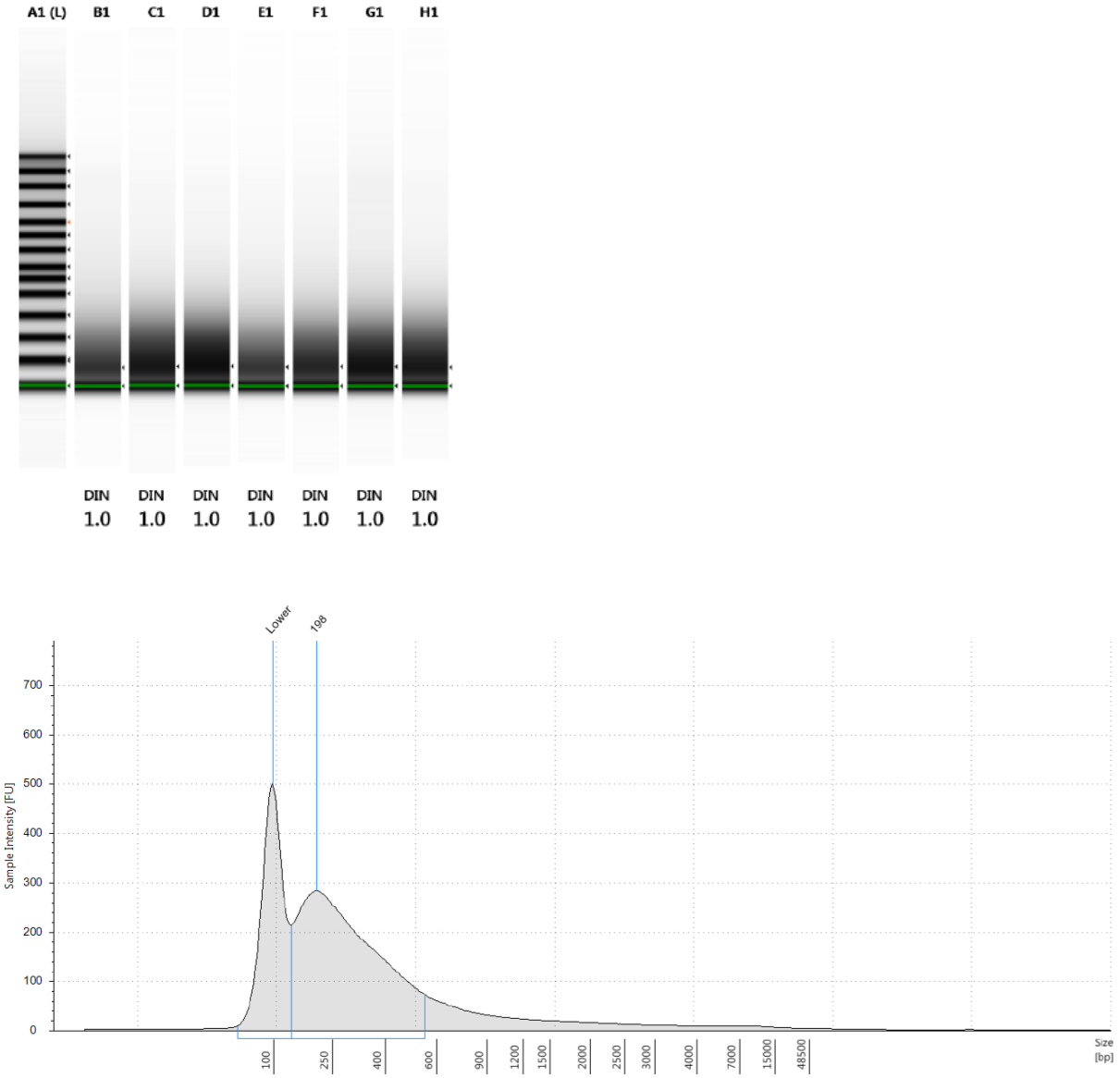



**Figure 4-2- Tapestation gel image for seven DNA extractions from mummified foetus 23982_R6 (top) and an example of one Tapestation electopherogram showing the distribution of DNA fragment sizes with a peak at 198 bp (bottom).**

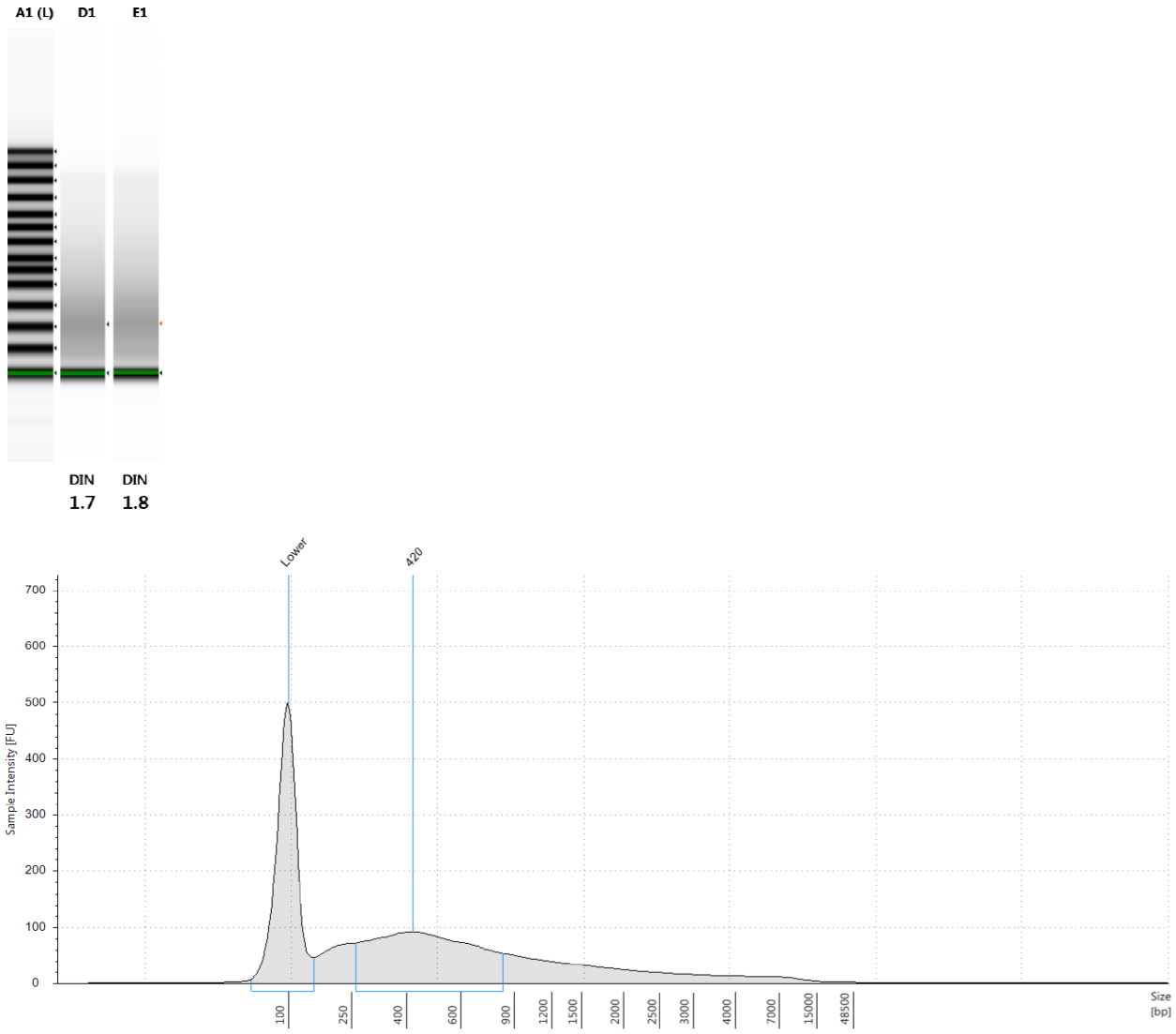



**Figure 4-3- Tapestation gel image for pooled DNA extractions from mummified foetus 23982_R6 in duplicate (top) and an example of one Tapestation electopherogram showing the distribution of DNA fragment sizes with a peak at 420bp(bottom).**

in table 4-1, these include details of quantity and quality after a clean-up step performed because the 260/280 ratio of CB was close to 2, suggesting RNA contamination.

### 4.3.3 Sequence quality control

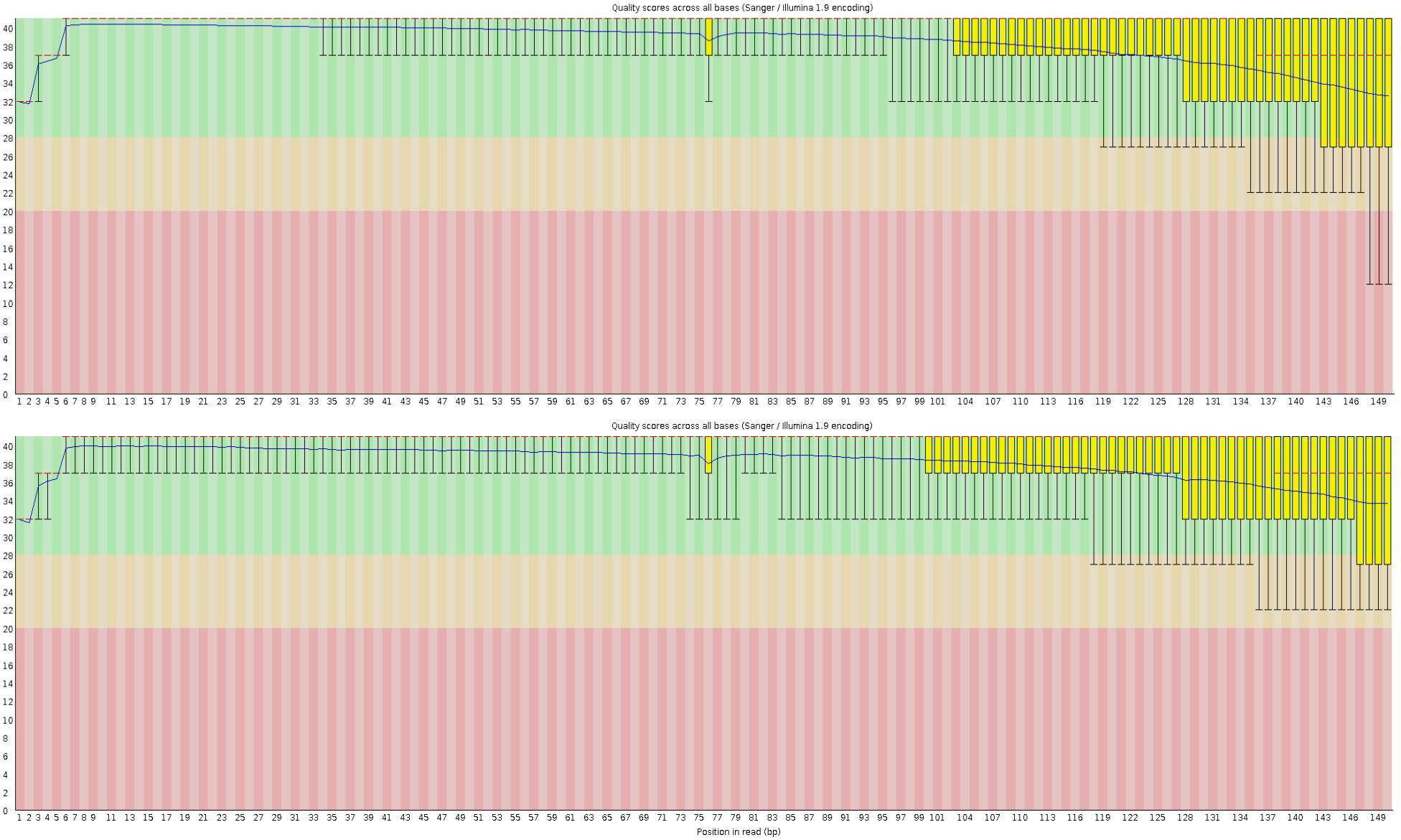
Initial QC with FastQC revealed no adapter content, all samples had similar good quality base calls (figure 4-4) and there was no apparent effect of degradation on the mummified foetus DNA. All three datasets had reduced quality scores towards the end of read 2, which is expected for Illumina data and in all cases this was much improved by trimming with sickle. Following alignment to Sscrofa10.2, samtools flagstat results (summarised in table 4-2) were good for all samples with little difference between 23982_R6, CB and 23982.

**Table 4-1- Table describing DNA quantity and quality of DNA extractions from CB and 23982 according to Nanodrop[1], Qubit[2], and Tapestation[3]**

| | CB | 23982 |
|---|---|---|
| **Initial DNA quantification[1]** | 696 ng/μl in 200ul | 582 ng/μl in 300ul |
| **Initial 260/280[1]** | 1.93 | 1.87 |
| **Initial 260/230[1]** | 2.34 | 2.44 |
| **Quantification post-cleanup[2]** | 153 ng/μl in 200ul | 569 ng/μl in 200ul |
| **Post-clean up DNA Integrity Number[3]** | 9.8 | 9.7 |

**Table 4-2- Summary of key Samtools flagstat results for sequencing data from the trio**

| | 23982_R6 | 23982 | CB |
|---|---|---|---|
| **Reads** | 873,387,682 | 1,015,450,594 | 995,551,926 |
| **Mapped** | 837,835,544 (95.93%) | 964,195,546 (94.95%) | 935,479,252 (93.97%) |
| **Properly paired** | 750,084,554 (88.56%) | 895,815,538 (89.52%) | 867,447,146 (88.39%) |
| **Singletons** | 9,978,170 (1.18%) | 12,531,984 (1.25%) | 14,249,752 (1.45%) |


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

202

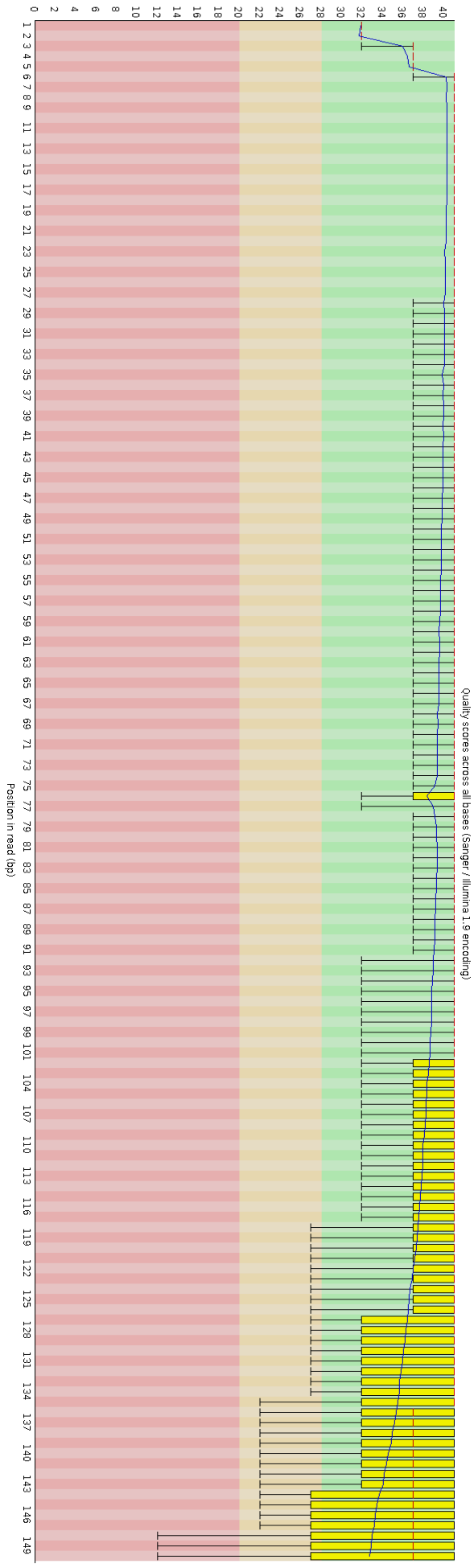



**Figure 4-4- FastQC plots for read 1 after trimming for 23982_R6 (previous page, left), CB (left) and 23982 (previous page, right)**

Genome coverage for 23982_R6 had a normal distribution over the autosomes according to BEDtools Genomecov (figure 4-5.A), and coverage over the X and Y chromosome was equal and roughly half that of the autosomes (figure 4-5.B-C) revealing the foetus to be male and suggesting little contamination from the dam's blood on the outer surface of the foetus during DNA extraction.

### 4.3.4 Variant calling, annotation and filtering

Variant calling on the trio identified 16,548,227 variants. In this dataset, there are 6,398,821 variants that appear in the homozygous state in CB and/or 23982. As the sire and dam are alive and capable of reproducing, these variants cannot be the sole cause of lethality in the embryo and were filtered out. Similarly, variants which are heterozygous or homozygous reference in 23982_R6, of which there are 8,178,882 and 3,899,814, respectively were filtered out. Finally, variants in which one or more individuals had no genotype called have been excluded. Further details of the number, consequences and filtering of the variants is summarised in figures 4-6 to 4-8. Filtering for variants that occur *de novo* in the homozygous state in 23982_R6 identified 6,026 variants, 6 of which were HIGH impact, however these are highly likely to be false positives and may relate to sequencing error or errors in the reference genome.

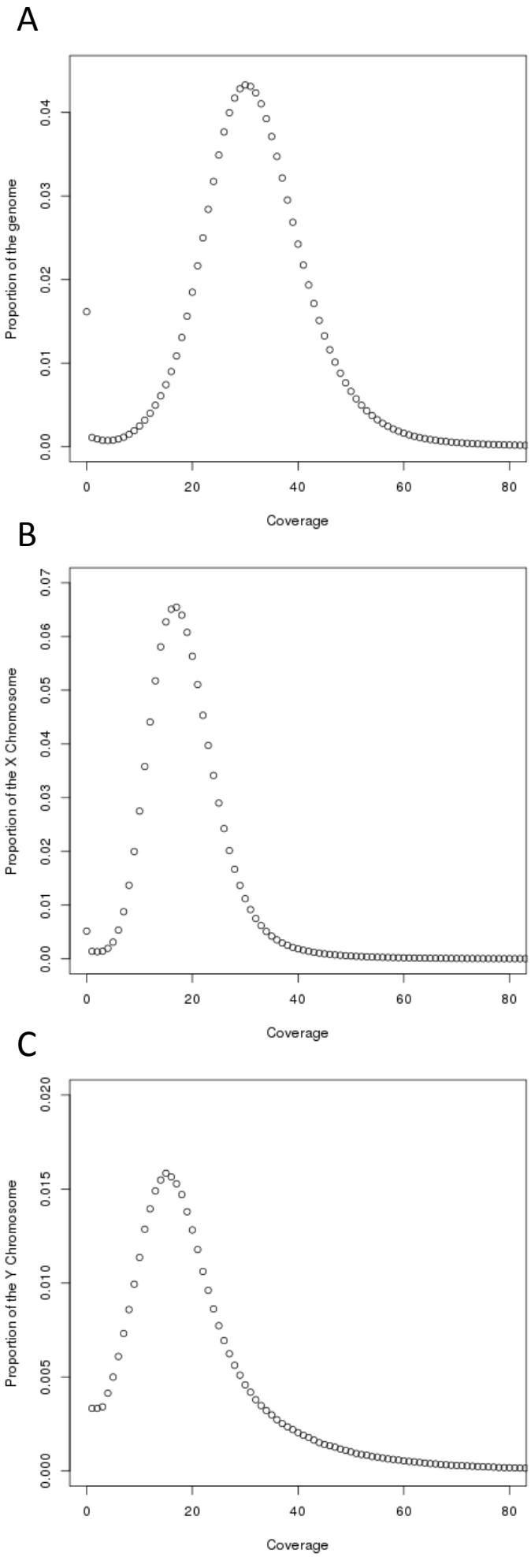



**Figure 4-5- (Right) Plots showing the average coverage of reads for 23982_R6 over the autosomes (A), the X chromosome (B) and the Y chromosome (C) based on results from BEDtools genomecov. Axes have been limited to exclude extreme coverage.**

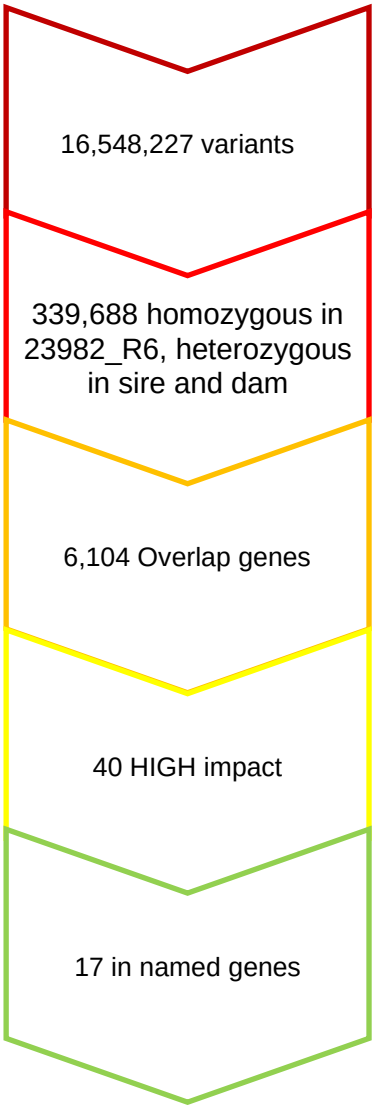



**Figure 4-6- Summary of filtering results from initial dataset to final candidates that were followed up on with a literature search**


**Figure 4-7- Pie chart of variant types from annotation of full set of variants found in the three samples**

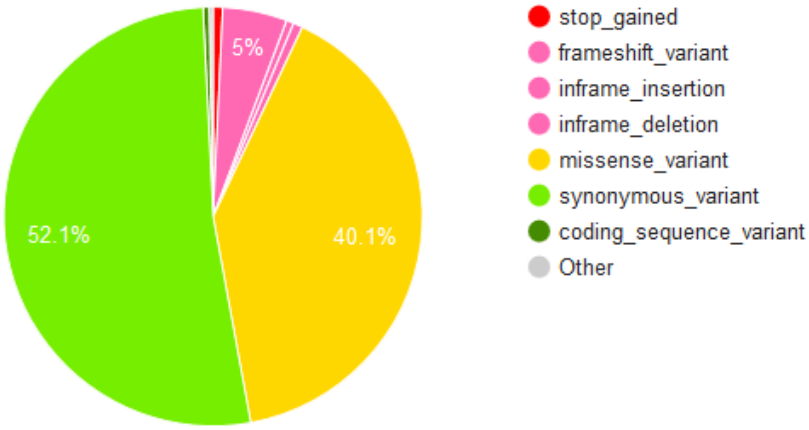



**Figure 4-8- Pie chart of annotated consequences of variants in full set of variants found in the three samples**

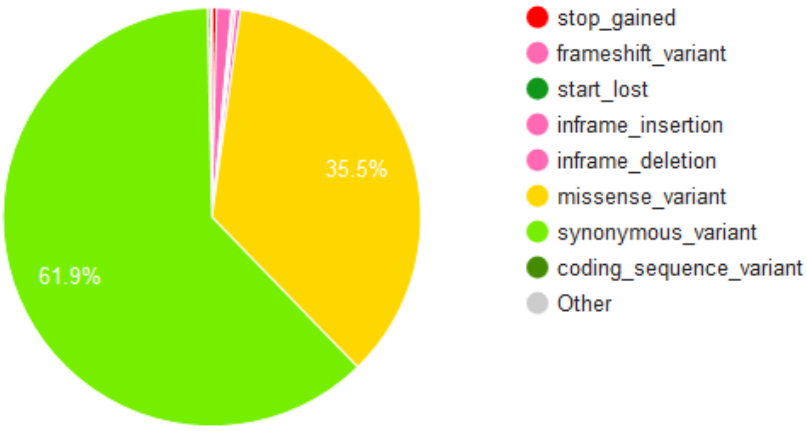



**Figure 4-9- Pie chart of annotated consequences of variants in genotype filtered set of variants from the three samples**

Following filtering for variants that are heterozygous in CB and 23982 and homozygous in 23982_R6, the proportion of variant types was similar, however the consequences tended to be of lower impact (figure 4-9), revealing a small pool of high impact candidates.

High impact variants that were heterozygous in CB and 23982 and homozygous in 23982_R6 are listed in table 4-3 and will be discussed further below.

**Table 4-3- Table of 40 candidate variants that are high impact, homozygous in the mummified foetus (23982_R6) and heterozygous in the sire (CB) and dam (23982). Final column refers to regions identified as low quality (LQ) or low coverage (LC) in chapter 2. Consequences are abbreviated: F=frameshift, SA=splice acceptor, SD=splice donor, SR=splice region, SL=stop lost, SG=stop gained, CDS=coding sequence variant**

| **Variant Position** | **Variant type** | **Gene name/*Ensembl ID*** | **Gene Name/*Human orthologue*** | **Consequence** | **LQ/LC?** |
|---|---|---|---|---|---|
| 1: 29114892 | SNP | NHSL1 | NHS Like 1 | SA | |
| 1: 31478180 | Indel | MTFR2 | Mitochondrial Fission Regulator 2 | F | |
| 1: 204304619 | Indel | CDKN3 | Cyclin Dependent Kinase Inhibitor 3 | F&SL | LC |
| 1: 293511480 | Indel | STOM | Stomatin | F | |
| 1: 304373070 | SNP | HMCN2 | Hemicentin 2 | SA | |
| 1: 305472018<br>1: 305472624 | Indel<br>SNP | *ENSSSCG00000027241* | Novel gene | F<br>SA | LQ |
| 2: 64446751 | SNP | *ENSSSCG00000013788* | Novel gene | SG | LC |
| 4: 99504244 | SNP | OR6K6 | Olfactory Receptor Family 6 Subfamily K Member 6 | SL | |
| 5: 88073588 | SNP | DEPDC4 | DEP Domain Containing 4 | SD | |
| 6: 125237842 | SNP | SURF6 | Surfeit 6 | SG | LQ |
| 6: 127958338 | SNP | SLC44A5 | Solute Carrier Family 44 Member 5 | SD | |
| 7: 14580986 | SNP | TPMT | Thiopurine methyltransferase | SG | |
| 7: 127767114 | SNP | BCL11B | B Cell CLL/Lymphoma 11B | SD | |
| 8: 81956756 | Indel | *ENSSSCG00000028236* | Novel gene | SA&CDS | |
| 9: 4272467<br>9: 4272469 | Indel<br>Indel | *ENSSSCG00000029114* | *Olfactory Receptor Family 56 Subfamily A Member 4 (OR56B4)* | F<br>F | |
| 9: 4344435<br>9: 4344437 | Indel<br>Indel | *ENSSSCG00000029794* | *Olfactory Receptor Family 56 Subfamily A Member 4 (OR56B4)* | F<br>F | |
| 9: 29559438 | SNP | *ENSSSCG00000025995* | Novel gene | SA | LQ |
| 9: 29728798<br>9: 29754894<br>9: 29754895<br>9: 29774982<br>9: 29777753 | SNP<br>Indel<br>SNP<br>Indel<br>SNP | CEP295 | Centrosomal Protein 295 | SG<br>SA<br>SA<br>F<br>SD | LQ |
| 9: 72610968 | SNP | SLC45A3 | Solute Carrier | SG | |

| | | | | | |
|---|---|---|---|---|---|
| | | | Family 45 Member 3 | | |
| 10: 15627047 | SNP | *ENSSSCG00000023390* | Novel gene | SL | |
| 10: 47679923 | Indel | *ENSSSCG00000027759* | Novel gene | F&SR | |
| 11: 78367604<br>11: 78367611 | Indel<br>Indel | *ENSSSCG00000009530* | *Coiled-Coil Domain Containing 168 (CCDC168)* | F<br>F | LQ |
| 12: 6321086 | Indel | *ENSSSCG00000025104* | ATP synthase subunit D | F | LC |
| 13: 25472029 | SNP | EXOG | Exo/Endonuclease G | SG | |
| 14: 6418550 | Indel | *ENSSSCG00000023463* | GDNF Family Receptor Alpha 2 | F | LQ |
| 14: 118747251 | SNP | *ENSSSCG00000010530* | Novel Gene | SG | LQ |
| 16: 71812695 | Indel | CYFIP2 | Cytoplasmic FMR1 Interacting Protein 2 | SD&CDS | |
| 17: 16110163<br>17: 16110167 | Indel<br>Indel | *ENSSSCG00000024507* | *Ankyrin Repeat Domain 26 (ANKRD26)* | F<br>F | LC |
| 17: 35256025 | Indel | NINL | Ninein Like | F | |
| 17: 57521610 | SNP | PTGIS | Prostaglandin I2 Synthase | SL | LC |
| 18: 6014060 | Indel | *ENSSSCG00000026023* | Protein Kinase AMP-Activated Non-Catalytic Subunit Gamma 2 | F | |

From this set of 40 variants, the rs332515507 SNP in the EXOG gene was prioritised as it had previously been identified as a high quality candidate variant (see chapter 3). Three additional variants were also prioritised based on their function via a review of the literature, these are the variants in SURF6, CYFIP2, and BCL11B. SURF6 is in a region previously identified as low quality in chapter 2, while this reduces the confidence in the annotation given to it, it has been taken forward due to the relevance of the gene to the phenotype. SURF6 is an important gene in implantation and the early embryo and knockouts in mice are lethal (Romanova et al., 2006), CYFIP2 knockouts are also lethal in mice (Kumar et al., 2013), and BCL11B has roles in the development of a number of important systems (Lennon et al., 2017). These are discussed further below.

The six high impact *de novo* variants that were homozygous in 23982_R6 were visualised with IGV and were all found to be false-positives. Two of the variants were large insertions that began at the edge of assembly gaps (e.g. figure 4-10) and likely represent missing sequence from Sscrofa10.2. Other variants in this set were in regions with poor mappability due to repeats and low-complexity (e.g. figure 4-11). Five were in LQLC regions identified in chapter 2, and the other one was 76 bases away from a region that was LQLC.

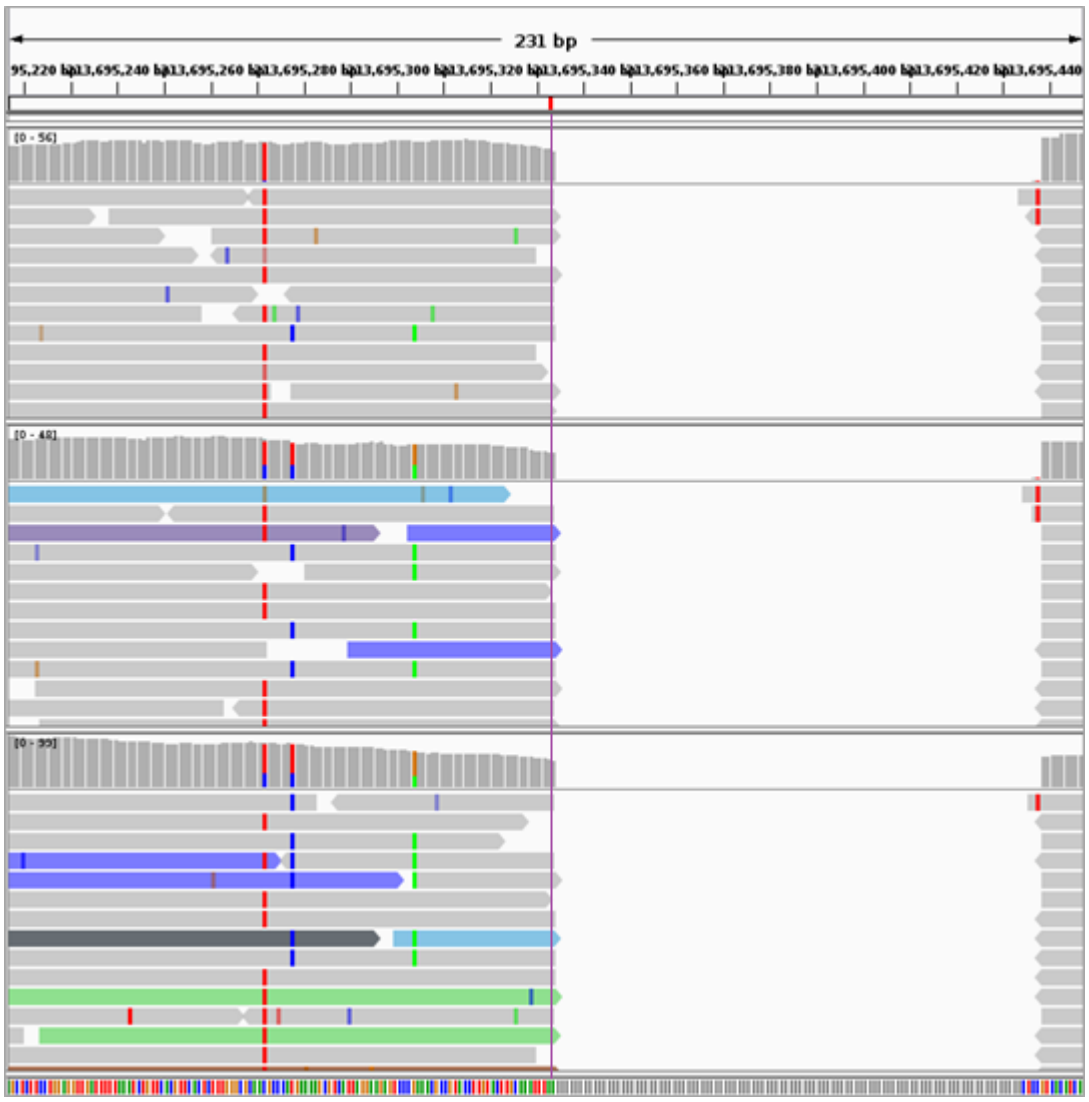



**Figure 4-10- Example of a variant called as homozygous in 23982_R6 and homozygouys reference in 23982 and CB. The variant is marked by a vertical purple line. There is a gap in the genome as represented by the grey bases on the lower track. The top track are reads from 23982_R6, the centre track are reads from 23982, and the bottom track are reads from CB.**


**Figure 4-11- Example of a variant called as homozygous in 23982_R6 and homozygous reference in 23982 and CB. The variant is shown as a 7 or 9 base deletion. There are mapping low quality reads in all three samples. The variant was called by GATK as a 9 base insertion of a T homopolymer. The low complexity context of this region can be seen in the lower track where there are CT repeats (blue and red) and a run of Ts (red). The top track are reads from 23982_R6, the centre track are reads from 23982, and the bottom track are reads from CB.**

No moderate or high impact variants that were heterozygous in 23982 and CB and homozygous in 23982_R6 were identified within the regions identified by Derks et al. (2017), with all variants being intergenic, intronic, or up- or downstream variants. It is possible one of these has a regulatory role rather than a direct effect on the protein, however further investigation of this is beyond the scope of this project. Additionally, no high impact variants were found in the unfiltered VCF in the BMPER gene, and one moderate impact variant was found. This variant was heterozygous in CB and 23982_R6 and homozygous for the reference allele in 23982 and so is unlikely to be relevant to the phenotype.

### 4.3.5 Genotyping

The candidate variants were genotyped in all of the remaining dams and healthy siblings and the remaining sire. The results of the genotyping for the siblings successfully genotyped are presented in table 4-4.

None of the adult animals were homozygous for any of the alt variants.

For the CYFIP2 indel, only 2 litters were from heterozygous crosses, the litter of sows 23982 and 23964, both sired by CB. Of the 13 seemingly healthy foetuses from 23982's litter, only one was homozygous for the deletion, with 8 homozygous for the reference allele and 4 heterozygous. Conversely, of the 13 seemingly healthy foetuses from 23964's litter only one was homozygous for the reference allele, with 7 heterozygous and 5 homozygous for the deletion.

**Table 4-4- Table of full set of genotyping results for the four main candidate variants**

| Gene | Variant ID | Homo ref | Het | Homo Alt | Number of individuals | Ref Allele % | Alt Allele % | Ref | Alt |
|---|---|---|---|---|---|---|---|---|---|
| CYFIP2 | - | 91 | 27 | 6 | 124 | 84.3 | 15.7 | TCCTC | - |
| BCL11B | rs344332274 | 37 | 65 | 19 | 121 | 57.4 | 42.6 | G | A |
| SURF6 | rs707030666 | 32 | 63 | 29 | 124 | 51.1 | 48.8 | A | T |
| EXOG | rs332515507 | 42 | 54 | 24 | 120 | 57.5 | 42.5 | C | T |

There were five litters from heterozygous crosses of the BCL11B variant. The average percent homozygous reference for these litters was 30.4%, the average percent heterozygous was 38.6% and the average percent homozygous alt was 28.3%.

There were three litters from heterozygous crosses of the SURF6 variant. The average percent homozygous reference for these litters was 20.5%, the average percent heterozygous was 45%, and the average percent homozygous alt was 32.6%.

There were three heterozygous crosses for the EXOG SNP. The average percent homozygous reference for these litters was 26.7%, the average percent heterozygous was 36.2% and the average percent homozygous alt was 37.1%.

## 4.4 Discussion

Embryonic losses are wasteful as the energy the dam puts into a foetus that fails to survive reduces the overall efficiency of the pregnancy. While the literature does suggest there may be genetic causes for litter size, stillbirth and mummification (e.g. Ghanem et al., 2006, Derks et al., 2017, Schneider et al., 2015), these are expected to be multifactorial phenotypes and not every stillborn or mummified foetus will have a genetic cause. Protein truncating variants are a good starting point when looking for causes of a phenotype, as any protein that is cut short has potential to be rendered non-functional and regulatory variants are much more difficult to annotate. However, protein truncating variants are actually common in healthy

individuals (Groenen et al., 2012, Lek et al., 2016, MacArthur et al., 2012, MacArthur and Tyler-Smith, 2010), and the function of genes must be used to further filter these. This filtering is limited by our understanding of the functions of genes, and redundancy of gene function in the genome.

While mummification is a dramatic phenotype, it is purely a consequence of the death of the foetus after the bones have started to form and there are thus many possible underlying causes. The death of a foetus can be from any number of genetic variants or environmental factors, and the timing of the death determines whether or not the foetus will mummify. Lethal variants may terminate a foetus before the bones form and the loss of the foetus may not even be noticed but simply identified as a sow producing a smaller litter size. As pigs are already under strong selection for litter size lethal variants are already under negative selection, additionally European pig breeds are expected to have fewer deleterious variants than most domestic species owing to their level of outbreeding (Makino et al., 2018), so it is perhaps not surprising that they are difficult to find in a commercial herd.

The work in chapter 3 employed stringent filtering, but trio sequencing offers a method of reducing a huge number of variants to a small number of candidates using a different kind of filtering that has less risk of discarding true-positive variant calls. A limitation of using trio sequencing without stringent filtering is that searching for variants such as *de novo* variants against a poor quality reference genome is likely to produce a list of false-positives. For phenotypes with a single causative variant against a high quality genome it has been demonstrated that trio sequencing can quickly

and correctly identify these variants (Meng et al., 2017a). In the case described here, focussing on variants that are homozygous in the offspring and heterozygous in the sire and dam reduced the number of variants from 16,548,227 to just 339,688. Not only was the total number of variants reduced, but the proportion of those variants that are predicted to be high impact was reduced. With as few as 40 high impact variants, these can easily be individually investigated for any relationship with the phenotype and inspected for reliability, moving the quality control of the variants from the first step to the last and avoiding the early loss of true positive variants through strict initial filtering. Ideally, all 40 of these variants would have been genotyped in the remaining individuals, however, with the number of samples involved this is prohibitively expensive.

Of the 40 variants in the filtered set, 18 were in unnamed genes, with Ensembl identifying most of them only as “novel gene”, with the exception of ENSSSCG00000025104 (ATP5H), ENSSSCG00000023463 (GFRA2) and ENSSSCG00000026023 (PRKAG2). While these genes have been identified through homology, they are incomplete, all three have poor percentage identity scores with the orthologous of genes in other species, and both ENSSSCG00000025104 and ENSSSCG00000023463 are in regions identified as LQLC in chapter 2. This renders any variant calls based on changes to codons in the exons of these genes unreliable. ATP5H is one of many ATP synthase genes, while knockout of some ATP synthase genes have been linked with embryonic death in mice (Vrbacký et al., 2016), this one specifically has not. This gene sits next to a reference gap and is likely

missing part of the gene making the consequence call less reliable, although it is a frameshift, so assuming it truly is in an exon of the gene, it may knock it out. While the quality of the individual variant is good, visualisation of the region on the gEVAL browser (Chow et al., 2016) reveals the surrounding region is abnormal (figure 4-12).

GFRA2 is a neurotrophic factor, and while knockout mice have a number of negative phenotypes (Rossi et al., 2003), they are viable. Again, this particular variant is supported by the reads, however the fragmented gene sits on a short contig between two reference gaps, and gEVAL visualisation of this region suggests a misassembled region (figure 4-13).

PRKAG2 is an enzyme that regulates glucose uptake and glycolysis and this is a highly conserved sequence. In humans only missense or in-frame indels have been described (Porto et al., 2016), and the ExAC database (Lek et al., 2016, Exome Aggregation Consortium (ExAC), 2014) estimates a LoF intolerance probability of 0.98. In this case CB and 23982 have only 8 reads each covering the variant loci, with 23982_R6 having 35 reads. Again, the region is next to a reference gap and appears to be of low quality (figure 4-14), although was not one of the regions identified in chapter 2 and would not have been filtered out if these regions had been used as a filter. The variant is however just 60 bases away from a region identified as LQLC, which can be appreciated in figure 4-14.

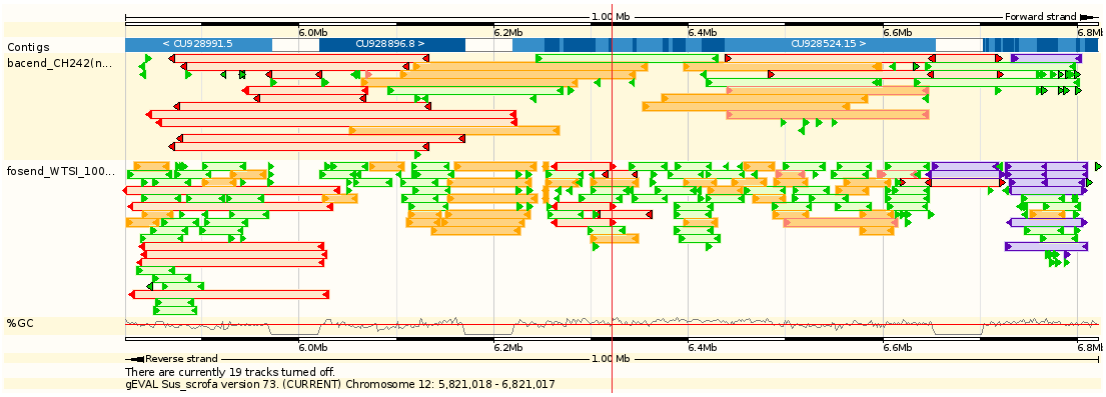



**Figure 4-12- Region surrounding a variant in the fragmented gene ENSSSCG00000025104 as visualised in the gEVAL browser. Vertical red line marks approximate location of the indel. The top track shows mapped ends of BAC ends from the library used in the assembly of the reference genome, bottom track shows mapped fosmid ends. Ends joined in red are in the incorrect orientation, in orange are incorrect insert size, in purple are single ends mapping multiple times, and in green are ends mapping with the expected orientation and insert size.**

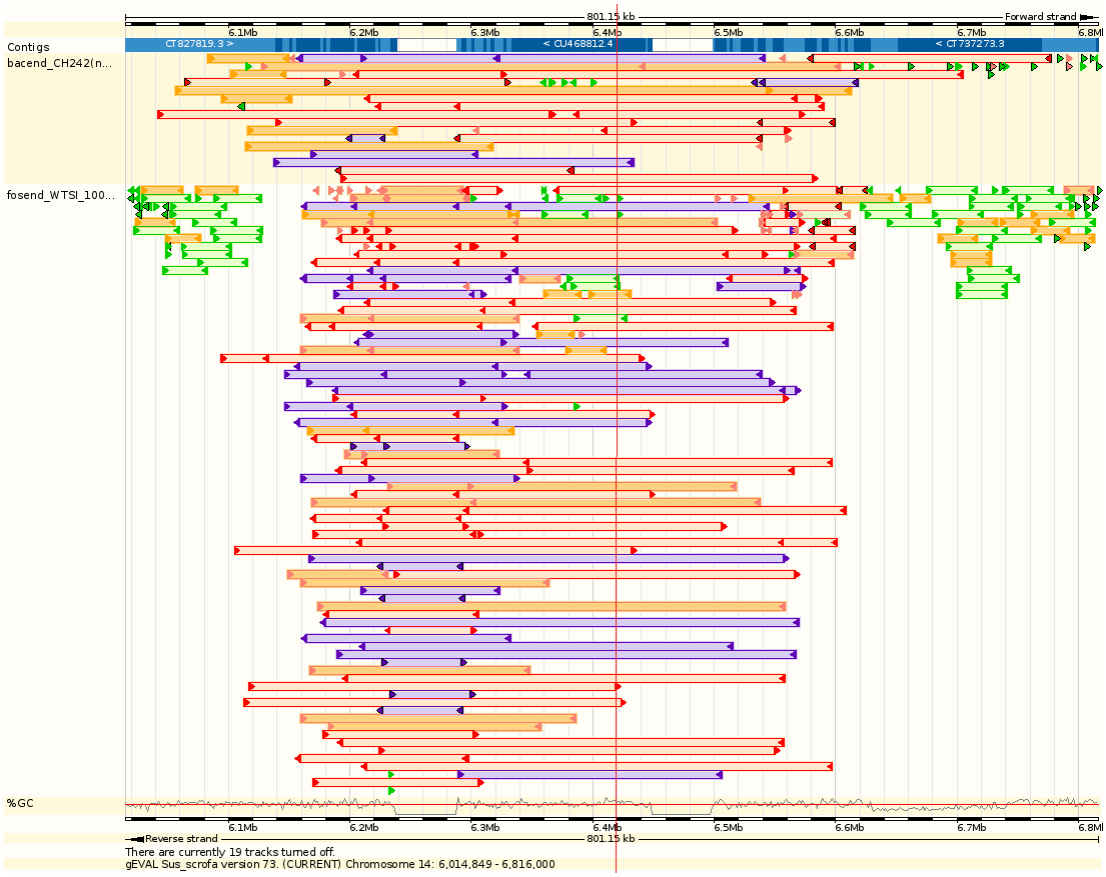



**Figure 4-13- Region surrounding a variant in the fragmented gene ENSSSCG00000023463 as visualised in the gEVAL browser. Vertical red line marks approximate location of the indel. The top track shows mapped ends of BAC ends from the library used in the assembly of the reference genome, bottom track shows mapped fosmid ends. Ends joined in red are in the incorrect orientation, in orange are incorrect insert size, in purple are single ends mapping multiple times, and in green are ends mapping with the expected orientation and insert size.**

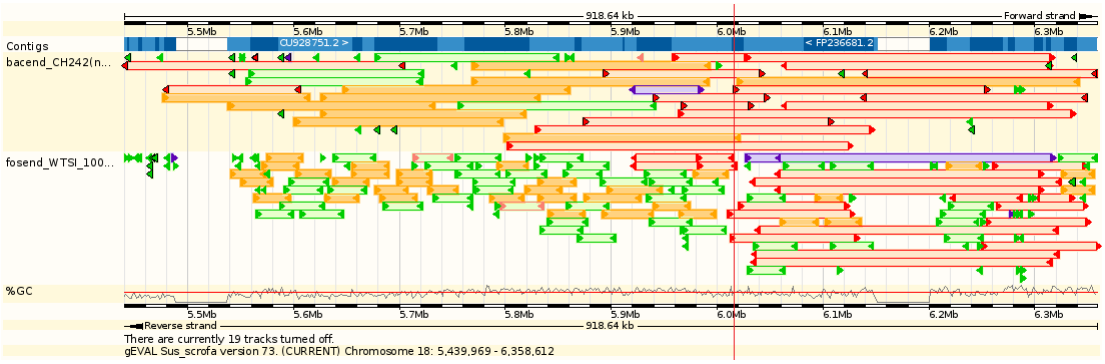



**Figure 4-14- Region surrounding a variant in the fragmented gene ENSSSCG00000026023 as visualised in the gEVAL browser. Vertical red line marks approximate location of the indel. The top track shows mapped ends of BAC ends from the library used in the assembly of the reference genome, bottom track shows mapped fosmid ends. Ends joined in red are in the incorrect orientation, in orange are incorrect insert size, in purple are single ends mapping multiple times, and in green are ends mapping with the expected orientation and insert size.**

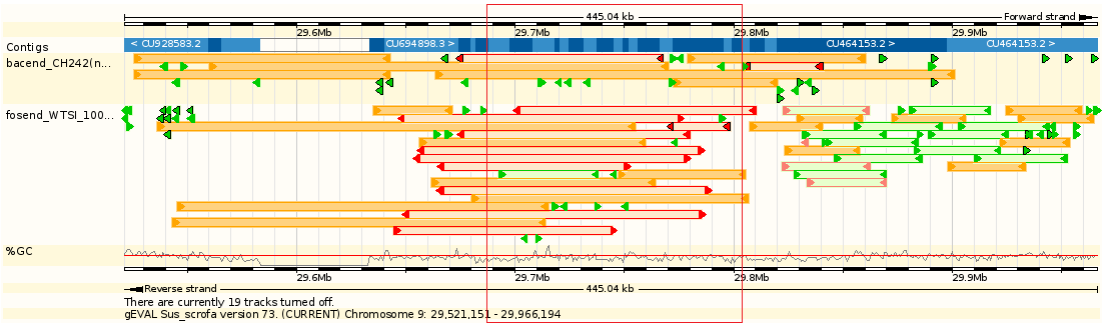



**Figure 4-15- Region surrounding the gene CEP295 as visualised in the gEVAL browser. The top track shows mapped ends of BAC ends from the library used in the assembly of the reference genome, bottom track shows mapped fosmid ends. Ends joined in red are in the incorrect orientation, in orange are incorrect insert size, and in green are ends mapping with the expected orientation and insert size. The blue track at the top switches between two shades to indicate contig boundaries.**

These fragmented genes demonstrate the importance of a contiguous and well annotated reference genome. While the first two genes described are unlikely to cause lethality when knocked out, the third is LoF intolerant. With regions of the reference genome being of poor quality, this variant and those in the completely unidentified genes that may be of interest must be overlooked to focus on the more high-confidence variant calls, despite their potential relevance to the phenotype.

CEP295 is another candidate in the list of 40, there are 5 different high impact variants called. CEP295 is a highly conserved centrosomal protein without which centrioles cannot form (Tsuchiya et al., 2016). It is extremely unlikely that there would be true knockout variants of CEP295 persisting in a population. This region, like the regions associated with the fragmented genes, appears to be poorly assembled with highly fragmented contigs (figure 4-15) and several of the variants in the gene are within the LQLC regions identified in chapter 2.

Of the remaining genes in the candidate list, OR6K6 and OR56B4 are olfactory receptors and are unlikely to cause a lethal phenotype. NHSL1, MTFR2, HMCN2, DEPDC4, SLC44A5, CCDC168 and SLC45A3 have insufficient published functional and knockout data available to associate them with the phenotype, and CDKN3 (Sun et al., 2016), STOM (Zhu et al., 1999), NINL (Adissu et al., 2014), PTGIS (Nakayama et al., 2002), ANKRD26 (Acs et al., 2015) and TPMT (Ramsey et al., 2014) have viable knockouts in other species.

This reduces the list to the four most likely candidates, two splice donor variants in CYFIP2 and BCL11B, respectively, and two stop gains in SURF6 and EXOG, respectively. For each of these the function is discussed below in addition to the estimated probability of loss of function (LoF) intolerance from the ExAC database (Exome Aggregation Consortium (ExAC), 2014). This number is based on the expected number of LoF variants for the size of the gene vs the number of observed variants. A LoF intolerance >0.9 is considered to be a gene highly intolerant of LoF variants.

CYFIP2 is a cytoplasmic protein involved in T-cell adhesion and homozygous knockout mice die at birth (Kumar et al., 2013) or shortly thereafter (Han et al., 2015) depending on the precise target of the knockout. The predicted probability of LoF intolerance in the ExAC database for this gene is 1. While this variant is a promising candidate, it is a splice variant, and therefore it is difficult to predict the exact effect it would have on the protein. Additionally, knockout of this gene is associated with post-natal death whereas the mummified foetus in question died sometime between day 30 and day 60 of gestation.

BCL11B is a transcription factor and has roles in the differentiation and development of neuronal subtypes, the immune system, integumentary system and cardiac system. Knockout of this gene is incompatible with extra-uterine life (Lennon et al., 2017, Arlotta et al.). Similarly to CYFIP2, knockout mice are born alive, but die shortly thereafter. The probability of LoF intolerance for this gene in ExAC is 0.93.

SURF6 regulates ribosomal biogenesis and is one of the earliest genes to be transcribed in the 1-cell embryo, reaching its peak expression in the 8-cell morula. In mouse knockouts development halts at the 8-cell stage, suggesting an essential role of SURF6 in pre-implantation development (Romanova et al., 2006). Just as the previous two variants cause death later in development than 23982_R6 died, this one would be more likely to cause death much sooner. ExAC predicts the probability of LoF intolerance to be 0, however there are no homozygous human knockouts observed in the database. This is the only one of the final prioritised variants to occur in the LQLC regions identified in chapter 2.

EXOG is involved in mitochondrial DNA repair, and while no mouse knockout has been produced, depletion of the protein has been associated with single strand breaks in mitochondrial DNA leading to apoptosis (Tann et al., 2011) and cardiomyocyte hypertrophy (Tigchelaar et al., 2016, Tigchelaar et al., 2014), and knockouts of genes involved in the same mitochondrial DNA repair process as EXOG are embryonic lethal (Ludwig et al., 1998, Xanthoudakis et al., 1996, Larsen et al., 2003, Puebla-Osorio et al., 2006). ExAC predicts the probability of LoF intolerance to be 0, however there are no homozygous knockouts observed in the database.

Aberrant splicing can be caused by variants that disrupt *cis*-acting elements of splice sites such as the 5` or 3` splice site, exon or intron splicing enhancers, and exon or intron splicing silencers. During splicing, the exon-intron boundary is identified in pre-mRNA by conserved sequences. Alternative splicing may occur to create different functional proteins by exon

skipping, the use of different 5` or 3` sites, mutually exclusive exons and intron retention. Mutations in splice site regions may disrupt normal splicing and have a deleterious effect on protein function (Jian et al., 2013). Disruptions to splice sites are known to have the potential to cause disease (e.g. Cartegni et al., 2006) with some work suggesting an estimated 15% of genetic diseases in humans can be attributed to them (Krawczak et al., 1992). Splice donor variants are difficult to interpret *in-silico* without additional data such as mRNA sequencing or experimental assays. Ensembl VEP predicts a high impact splice donor/acceptor variant if the first or last two bases of an intron are modified, or a low impact splice region variant if there is a modification within a few bases of the splice site. While *in-silico* tools have been designed for predicting the location of splice sites and the impact of splice variants (e.g. Rogozin and Milanesi, 1997, Pertea et al., 2001, Nalla and Rogan, 2005, Divina et al., 2009, Lim and Fairbrother, 2012), these have not been rigorously tested and must be followed up with expensive *in-vitro* work.

All of the variants genotyped were found in the homozygous form in seemingly healthy foetuses. However, none of them were found homozygous in any of the adults and it is not known if the homozygous individuals would have survived until birth. The variant in EXOG was also a candidate in the previous chapter and in that case there was an adult found to be homozygous, however it was unknown if the variant might make an individual less likely to survive. From the results seen here, it seems unlikely that the EXOG SNP has an effect on survivability, in litters resulting from

heterozygous crosses the homozygous stop-gain was actually slightly more common than the homozygous reference and the allele representation across the whole dataset was roughly even. Similarly, the alleles of the variants in BCL11B and SURF6 were evenly represented and the genotype distribution of the offspring of heterozygous crosses was not different than expected. The deletion in the splice site of CYFIP2 was underrepresented in the population, with an allele frequency of ~16% and only 6 individuals homozygous, although it was only present in three of the adults so a low allele frequency for the whole set is not unexpected. For the litter the mummified foetus originated from there was only one healthy sibling homozygous for the deletion making it a promising candidate. However, in the only other litter from a heterozygous cross the opposite was true with only one individual homozygous for the reference sequence, the inheritance pattern of the variant for these two litters is unusual but not impossible.

The list of 40 candidates includes only those that Ensembl VEP has annotated as high impact, while these are the variants most likely to affect the function of the protein, there are many examples of missense variants causing detrimental phenotypes (Stefl et al., 2013). The list of missense variants that were homozygous in the foetus and heterozygous in the sire and dam is much longer. One way to reduce this number would be to also whole genome sequence one or more siblings of the affected individual. This is a method often applied in paediatric diagnosis and allows the elimination of variants that match the genotype of the healthy sibling or retain variants that occur in multiple affected individuals from the same family. Additionally,

sequencing of a cohort of trios produces a large pool of sequence data of healthy, reproductively capable adults allowing for further filtering of the variants in the affected individuals. This would also allow for homozygous variants that occur with high frequency in the affected individuals to be prioritised without the variant necessarily having to be causative in all of the cases.

Large structural variants have the potential to cause disease phenotypes (Stankiewicz and Lupski, 2010). Finding structural variants in this trio is complicated by the poor quality of the reference genome Sscrofa10.2, and by the use of short-reads which are difficult to call large structural variants from and produce a large number of false-positives and false-negatives (Smith and Yun, 2017, Noll et al., 2016). Ideally, structural variants can be called from long-read sequencing data (Hampton et al., 2017, Merker et al., 2017) aligned to a high-quality reference genome. In this particular case, the highly fragmented DNA from the foetus would make this very difficult. It is possible that with repeated DNA extractions and the use of technology that selects long fragments such as pulse-field gel electrophoresis that it could be done, however the foetus is very small and obtaining sufficient DNA for this would be unlikely and impractical. Additionally, while contamination appears to be minimal in the sample sequenced, selecting for longer fragments may bias the sample towards DNA from maternal contamination if it is present. Alternative methods of DNA extraction could also be explored, in this case a method using spin columns was used, which may further fragment fragile DNA. For trio sequencing for other phenotypes with less degraded DNA,

long-reads would be ideal for searching for structural variants, however they should not be used for SNP and indel discovery due to their high error rates.

Deleterious variants tend to accumulate with increased inbreeding (Paige, 2010) and so the likelihood of lethal variants persisting in a population will depend on the size of the population and the level of inbreeding. In smaller populations deleterious alleles may increase through genetic drift (Paige, 2010, Star and Spencer, 2013). The population in this study is a large one and is under strong selection for reproductive traits such as litter size, as the observed maternal phenotype in cases of early embryonic death is reduced litter size, many variants that cause this will be being selected against already both through genomic selection and traditional selection, however purging rare and recessive detrimental variants entirely from a population is challenging (Derks et al., 2017). The populations discussed in Derks et al. (2017) are separate populations from those in this study, the lethal variants present in these populations are likely to differ from those in the current population due to differences in the history of the populations and the differing occurrences of mutations and inbreeding in those populations. The candidates identified in this study are not found in the candidate regions identified by Derks et al. (2017). The mummified foetus sequenced here may have died from either a genetic variant in another region, a variant with some regulatory impact, or an environmental factor. However, if a larger number of trios from the same population were sequenced it may be possible to identify genetic variants within the population that are associated with mummification. Furthermore, whereas the study using the 80K SNP chip was able to identify

associated regions, trio WGS has the potential to identify the causative variant itself.

This chapter has used trio sequencing of a mummified foetus, its sire and its dam to search for candidate variants that may cause embryonic lethality. The method allowed for rapid filtering of a large set of variants and identified a number of candidates. While several of these candidates appeared promising, further investigation of these revealed that none of them was likely to be the sole cause of lethality in the foetus. While 23982_R6 was selected from a litter with roughly a quarter of the foetuses deceased, this does not necessarily mean there is a genetic component in this case. Importantly, the ease by which a dataset of millions of variants can be filtered down to a manageable set of candidates makes the methods employed here attractive for identifying candidates in other trios both for mummification and for other traits of interest.

The lack of a clear candidate from this work may be in part due to the nature of the phenotype, in order to overcome the multifactorial nature of the phenotype, further work looking at a cohort of trios would allow for identification of variants that are more common in the foetuses, though not necessarily present in all of them. This would also include sequencing of a larger number of healthy adults with which to assist in filtering, which may reduce the candidate pool further and allow for further investigation of variants that are not predicted to be high impact but may be missense variants, or have some impact on regulation of a gene important for embryo

survival. Unfortunately, the cost of this is still likely to be prohibitively expensive.

Many aspects of the trio sequencing method and others that employ strict filtering are designed overcome the issue of a poor-quality reference genome, however these can only go so far. While trio sequencing is a method that avoids the strict filtering required for many other NGS methods, the results presented here are still full of false-positives, plagued by incomplete annotations and challenging for structural variant calling. Many of these issues can only be overcome through aligning NGS data to a higher-quality reference genome.

# CHAPTER 5: ASSEMBLY OF A NEW PIG REFERENCE GENOME, SSCROFA11.1

---

*"The foremost cartographers of the land have prepared this for you; it's a map of the area that you'll be traversing."*
*[Blackadder opens it up and sees it is blank]*
*"They'll be very grateful if you could just fill it in as you go along."*
-Blackadder II

## 5.1 Introduction

The first methods for sequencing DNA were established in the 1970s with the introduction of Sanger sequencing (Sanger et al., 1977, Sanger and Coulson, 1975) and Maxam & Gilbert sequencing (Maxam and Gilbert, 1977). These methods were labour intensive, time consuming, and could sequence only short sections of DNA, but nonetheless were a huge breakthrough paving the way to our current understanding of genetics and genomics. Since DNA sequencing became available, the challenge of sequencing an entire genome of a complex organism has been an important goal, but for a long time was a huge undertaking. Soon after sequencing methods were established, the assembly of DNA using computational methods based on overlapping sequence was proposed by Staden (1979).

Mammals, including humans, mice and pigs, have large, complex genomes with high proportions of repetitive content. In the early days of DNA sequencing, the sequencing of a genome alone was considered a mammoth task, but the added problem of reassembling the sequences only added to the challenge. Initial attempts to understand the information in the human genome focussed on only the coding sequence, sequencing cDNA clones from transcribed sequences. However, it was soon recognised that this method could not capture regulatory elements and was insufficient and so plans were made to sequence and assemble an entire human genome (Dulbecco, 1986, National Research Council, 1988, Chen, 1989).

While unique sequences are relatively simple to overlap and combine into contiguous sequence (contigs), the major challenge comes from the repeat regions. In large, complex genomes such as the pig or human, repetitive content can be ~40% of the genome (de Koning et al., 2011). These repeats can only be properly assembled if the length of the reads is sufficient to span the repeats and reach into unique sequences either side to give them region-specific context (see figure 1-3). Additionally, repeat sequences and low complexity regions often appear in multiple locations throughout the genome, making the “anchor” sequence either side of the repeat essential for reducing gaps and fragmentation in the final genome assembly. In the early days of genome sequencing and assembly, there were no technologies capable of producing reads sufficiently long to handle this problem. The solution was to break the genome into smaller fragments, sequence and assemble those smaller fragments without the complications introduced by similar regions in other parts of the genome, and subsequently reassemble these fragments into a whole genome (see figure 1-2). This method was first implemented to assemble the genome of the nematode and model organism, *Caenorhabditis elegans* (Sulston et al., 1992), and was subsequently employed for the human genome (International Human Genome Sequencing, 2001, International Human Genome Sequencing Consortium, 2004) and the pig genome (Groenen et al., 2012, Schook et al., 2005) based on a physical map by Humphray et al. (2007). The method involved fragmenting the genome and creating a library of bacterial clones which harbour a fairly large fragment of the target genome (see figure 1-1). Clones from these libraries were

selected for sequencing based on their position in a physical map of the pig genome. The physical map has been constructed based on similarities in patterns (fingerprints) of *Hin*dIII digest fragments of each Bacterial Artitficial Chromosome (BAC) clone (Humphray et al., 2007). BAC clones were selected to represent a minimum tiling path, i.e. the smallest number of cloned sequences that can be sequenced to cover the entire genome. Each of these was then sequenced using a shotgun approach and assembled individually, simplifying the assembly problem. A similar strategy has recently been employed Zheng et al. (2016) and is now available from 10X Genomics in which long fragments (~40-60kb) are isolated in a droplet and fragmented and barcoded within the droplet before high throughput Illumina sequencing so that they can be associated with their genomic location later. However, this is primarily advertised and more commonly used as a method for phasing and scaffolding purposes rather than *de novo* assembly.

Ideally, the problem could be made easier if the reads were longer to begin with. Over the course of the Human Genome Project, the read lengths attainable from Sanger sequencing increased from 300-400 bp to around 700-900bp. Using these longer WGS Sanger sequences, whole-genome shotgun sequencing was used to produce genome assemblies for the dog (Lindblad-Toh et al., 2005) and the horse (Wade et al., 2009). However, the real breakthrough for the long-read approach to genome assembly has come with the introduction of long read sequencing technologies from Pacific Biosciences single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT), both capable of sequencing tens of

kiliobases of DNA per read. Using specific methods of DNA extraction and library preparation (Quick, 2018), ONT can even deliver so-called “ultra-long” reads greater than 2 Mb (Payne et al., 2018). Very long repetitive regions such as centromeres are still difficult to span, however a combination approach of BACs tiling across a human centromere sequenced with ONT have produced a closed human Y centromere (Jain et al., 2018b).

SMRT sequencing allows for high-throughput sequencing with no polymerase chain reaction (PCR) step to introduce GC-bias and, importantly PacBio SMRT sequencing can produce a read N50 of around 20kb. These long-reads are capable of spanning repetitive regions and incorporating them into longer, more unique sequence contexts, allowing them to be placed in the correct genomic position during assembly (Chin et al., 2013, Rhoads and Au, 2015). A limitation of PacBio sequencing is its high error rate (11-15%; Korlach, 2015, Rhoads and Au, 2015, Watson, 2018) which largely consists of indels, however these errors are mostly not context specific and tools exist to remove these using hybrid methods (Koren et al., 2012, Salmela and Rivals, 2014), or with sufficient coverage, by taking the consensus of the PacBio reads (Quiver (Chin et al., 2013) and Arrow from PacBio), or by use of more base-accurate sequencing technologies such as Illumina using tools such as Pilon (Walker et al., 2014).

Reference genome sequences that are moderately accurate and contain the majority of the species’ genes are often sufficient for QTL mapping and GWAS. However, as the community increasingly moves into the realm of exome and genome sequencing to identify causative variants the accuracy of

the reference genome becomes increasingly important. The analyses discussed so far in this thesis have one major conclusion in common: the published draft pig reference genome is insufficiently accurate (Groenen et al., 2012, Warr et al., 2015b). In the preceding chapters and in other studies an abundance of false-positives are produced, in studies employing GWAS wrongly positioned contigs obscure associations (van Son et al., 2017, Yang et al., 2013), and in structural variant analyses the many contig breaks, wrongly placed contigs and gaps in the genome will break these up and introduce false-positives (Warr et al., 2015b).

Here I present the assembly of a pig genome using 65X genome coverage of PacBio sequence reads from the same individual that was sequenced for the assembly of Sscrofa10.2, a Duroc sow named T. J. Tabasco (or Duroc 2-14). By using the same individual the data available from previous work can be used to quality control, error correct and compare the two assemblies of the same genome and visualize the correction of misassemblies from Sscrofa10.2. Some low-confidence regions are expected to remain, these will mostly relate to regions with low-mappability or extreme GC-contents which may be prone to false-positives when aligning short-read sequencing data to the assembly.

## 5.2 Methods

The author of this thesis was responsible for the assembly from the initial contigs produced by Falcon to submission to NCBI. All analyses after the

Falcon assembly reported here were carried out by the author of this thesis unless stated otherwise.

### 5.2.1 Sample and sequencing

Extraction of high molecular weight DNA was done by Laurie Rund at the University of Illinois. DNA was extracted from Duroc 2-14 cultured fibroblast cells passage 16-18 using the Qiagen Blood & Cell Culture DNA Maxi Kit, producing 139.15 µg DNA from three extractions. High molecular weight DNA was obtained through size selection using the BluePippin (Sage Science; MA, USA) device. The high molecular weight DNA was sequenced by PacBio using their long read sequencing technology. Libraries for SMRT sequencing were prepared and sequenced as described previously (Pendleton et al., 2015) using P6-C4 chemistry on the PacBio RSII using a total of 213 SMRT cells.

### 5.2.2 Assembly

Unless stated otherwise tools were run with default parameters and standard protocols.

#### *5.2.2.1 Initial contigs*

Contigs were assembled from reads >13 kb using the Falcon (v0.4.0) assembly pipeline following the standard protocol by Richard Hall at PacBio. Falcon produces “haploid” and “alternative” contigs, the former are the main assembled contigs, with the alternative generally being shorter contigs representing some of the alternative haplotypes of the contigs present in the haploid set. Quiver v. 2.3.0 (Chin et al., 2013) was used to correct the haploid

and alternative contigs using the PacBio reads. Only the haploid contigs were used in the next stages of the assembly, of which there were 3,206.

### *5.2.2.2 Contig quality control and splitting*

Paired-end Illumina reads from the same individual (http://www.ebi.ac.uk/ena/data/view/PRJEB9115) were mapped to the 3,206 haploid contigs and assessed for structural abnormalities using similar methods to those described previously (Chapter 2; Warr et al., 2015b). Briefly, 1000 bp windows across the contigs were assessed for levels for abnormal mapping including high GC-normalized coverage, improper pairing and unexpected insert sizes. Additionally BAC end sequences (BES; CHORI-242 library; Humphray et al., 2007) and fosmids (WTSI_1005 library: https://www.ncbi.nlm.nih.gov/clone/library/genomic/234/; ENA accession:HE000001 – HE565349; Skinner et al., 2016) from the same individual were mapped to the contigs and regions with multiple occurrences of incorrect orientation were examined manually in the Integrative Genomics Viewer (IGV; v.2.3.88; Robinson et al., 2011). In the cases where there was consistent evidence of structural disagreement between the contigs and the Illumina reads, BAC ends and fosmids, the contigs were split or trimmed.

### *5.2.2.3 Scaffolding*

In order to establish initial scaffolds, the contigs were mapped to Sscrofa10.2 using Nucmer (v3.23; Kurtz et al., 2004). The positioning of the contigs was determined by using the longest ascending subset of mapping locations using the show-coords tool from Mummer with the -g flag. Contigs with an

identity below 95% were excluded. Contigs that mapped to regions substantially larger (>180%) or smaller (<10%) than the contig size were excluded. These tolerances were intentionally lenient due to the inflated gap sizes (e.g. including 50 kb runs of Ns between scaffolds as required by the NCBI submission system in 2011) and highly fragmented nature of certain regions of Sscrofa10.2. Adjacent contigs were merged into a single fasta entry with Ns representing gaps between them. Gaps were estimated from the distance between the mapping locations against Sscrofa10.2, with an upper limit of 50 kb. Several of the remaining contigs were placed by identifying their longest alignment position, if this alignment was more than 50% the length of the contig and the aligned region overlapped with a gap with an identity >90% they were placed in the gaps leaving 25 bp gaps either side.

#### *5.2.2.4 Gap filling*

PBJelly (PBSuite v.15.8.24; English et al., 2012) was used with the 65X raw PacBio reads to fill the gaps in the scaffolds, default parameters were used for all stages except the assembly stage where max wiggle (-w) was set to 100 kb and max trim (-t) was set to 1000 bp. These parameters were changed to account for the extremely inaccurate gap sizes and missing sequence in Sscrofa10.2 that will have influenced the estimated gap sizes in this assembly and to allow overlapping contigs to be closed. Following initial gap filling, PBJelly was rerun on the fasta output from the first round, with the unused contigs from the Falcon output included to allow extension of the scaffolds. These contigs had been excluded initially to reduce secondary

mapping between scaffolds and unplaced contigs. PBJelly is able to add contigs to the end of scaffolds, but not place whole contigs in gaps, so the initial mapping of contigs to scaffolds was examined to find if any of the contigs that had been excluded in this stage due to overlap with existing contigs might fill the gaps. Contigs were placed on a case-by-case basis if there was evidence of overlap with placed sequence on both sides of the gap, if the initial contig quality control was relatively good, and if placement was well supported by BAC end mapping. Additionally, BACs for which the end sequences mapped to adjacent contigs and for which finished quality sequence was publically available, were aligned and the gap filled and placed following the same restrictions as the unplaced Falcon contigs. On completion of these gap-filling procedures 108 gaps remained. Estimation of the size of the remaining gaps was based on BAC end mapping, using the known median insert size of the CHORI-242 library (see https://bacpacresources.org). Any gaps estimated to be <100 bp were sized at 100 bp and unspanned gaps were sized at 50 kb. Following gap filling the scaffolds were aligned to Sscrofa10.2 to look for any major discrepancies between the two assemblies at the point where gaps had been filled. Some of the gaps filled appeared very repetitive. The previously described BACs and fosmids were aligned to the scaffolds and were found to support the gap joins.

#### *5.2.2.5 Additional sequencing of targeted BACs to fill gaps*

Five BACs from the CHORI-242 library were selected for further sequencing (CH242-188M9 (SSC16); CH242-323K10 (SSC18); CH242-284F8 (SSC18);

CH242-61K12 (SSC1); CH242-168C15 (SSC12)) based on BAC ends mapping either side of gaps. The BAC clones were obtained from BACPAC Resources (Children's Hospital Oakland Research Institute, CA, USA; https://bacpacresources.org) and DNA was extracted using the Epicentre BACMAX DNA purification kit following manufacturer's instructions. The BAC DNA was sequenced using Oxford Nanopore Technologies' MinION sequencer using a barcoded 2D library following the discontinued protocol SQK-LSK208 on an R9 flow cell using MinKNOW v1.0.5. This was done by the author of this thesis at the Porecamp 2016 workshop. Sequences were assembled using Canu (v.1.5; Koren et al., 2017) with default settings and each produced a single contig. The BAC vector sequences were removed from the contigs, the contigs were mapped to the assembly initially with Nucmer (v3.23; Kurtz et al., 2004) to confirm they mapped to the expected locations, with exact positions for placement determined by BWA mem (v.0.7.15; Li, 2013).

#### *5.2.2.6 Polishing*

Owing to the high error rate of PacBio sequencing, additional polishing with both the original 65X PacBio data and with Illumina data was carried out to reduce the error rate. Error correction was done using Arrow from the GenomicConsensus suite (https://github.com/PacificBiosciences/GenomicConsensus) using the PacBio data. This was followed by Pilon (v1.22; Walker et al., 2014) with fixlist restricted to "bases", but otherwise using default parameters and the paired-

end Illumina short read data used in chapter 2 that provided 50X genome coverage.

### 5.2.3 RepeatMasker

RepeatMasker v.4.0.6 (Smit, 2013-2015) was used to identify repeats in the genome using parameter --species pig.

### 5.2.4 Quality control of the assembly

Following assembly, the methods described in chapter 2 were carried out to assess the quality of the assembly. Additionally, markers from a radiation hybrid (RH) map (Servin et al., 2012) were aligned to the genome to assess the agreement between the genome and the expected order of markers. SNPs and indels were called by mapping the same Illumina data used previously to the assembly using BWA mem (v.0.7.15) and calling variants using GATK HaplotypeCaller as described previously to assess the error rate.

### 5.2.5 Submission to NCBI

Following polishing, the assembly was submitted to NCBI as Sscrofa11 (GenBank accession: GCA_000003025.5). As this assembly was based on a female pig, a Y chromosome assembly (GenBank accession: GCA_900119615.2) was sourced from Skinner et al. (2016) and added to Sscrofa11, the combined assembly is named Sscrofa11.1 (GenBank accession: GCA_000003025.6).

### 5.2.6 External assessments of quality of the assembly

The assembly was run through the gEVAL pipeline (Chow et al., 2016) by William Chow and Kerstin Howe from the Wellcome Sanger Institute. The assembly was also analysed using the Cogent pipeline (https://github.com/Magdoll/Cogent) by Elizabeth Tseng from PacBio. The author of this thesis was not involved in the design or running of these pipelines, but the results will be discussed as they reflect the quality of the assembly.

The gEVAL browser will also be used to visualise some notable differences between Sscrofa10.2 and Sscrofa11.1.

### 5.2.7 Annotation

Annotation of the genome was carried out by Ensembl and NCBI. The author of this thesis was not involved in the annotation process, but the results will be discussed as the completeness of the annotations reflect the quality of the assembly.

### 5.2.8 Detection of remaining errors post-annotation

Following annotation, some genes were found to be truncated during unrelated analyses using this genome assembly as a reference. Additionally, other researchers reported indels in some genes of interest. Despite polishing, PacBio genome assemblies often have more indels remaining than genomes assembled from more accurate technologies (Watson, 2018, Jain et al., 2018a, Watson and Warr, 2019). Pilon recommends 100X Illumina coverage or greater, in this case only 50X was available and so this

assembly may have a greater indel rate than some other assemblies that use this method. The main variants observed in these truncated genes appeared to be two false indels a short number of bases apart that occur on opposite haplotypes. When Illumina data from the reference animal was aligned to these regions, GATK either called them as two heterozygous indels a short distance apart, or only called one of the two indels. Through visualisation with IGV and comparisons with results from Sscrofa10.2 it became clear that these were heterozygous SNPs whose reference base was absent from the assembly. A problem with PacBio's Quiver and Arrow tools has been reported informally, and recently addressed by PacBio (Drake, 2018) in which the tools do not correctly handle heterozygous SNPs in the reference individual, interpreting them as two false indels in the same location and removing the reference allele base from the reference (figure 5-1). It was believed Pilon would be able to resolve these using Illumina data, however it appears in certain sequence contexts that this is not the case. To assess the extent of this problem, further analyses were undertaken to identify remaining indels of this sort in the dataset.

While GATK did not consistently identify these variants, the Platypus haplotype caller (Rimmer et al., 2014) appears to call them reliably. In the output VCF from Platypus, different haplotypes are annotated differently in the genotype field, with "0/1" representing one haplotype and "1/0" representing the other which made these false indels easier to identify. Platypus (v0.8.1) was used to call variants in Illumina data aligned to Sscrofa11 using BWA mem (v.0.7.15). Custom scripts were used to identify

pairs of indels occurring between 1 and 10 bases apart on opposite haplotypes and of the two indels the one with the highest quality score was selected to be corrected. Through visualisations of 50 of these at random in IGV, a cut off value was selected for the Platypus quality score. All of the variants inspected with a quality value of >130 appeared reliable and this was selected as a cut off value. Some of these deleted SNPs may remain after correction in low coverage regions.

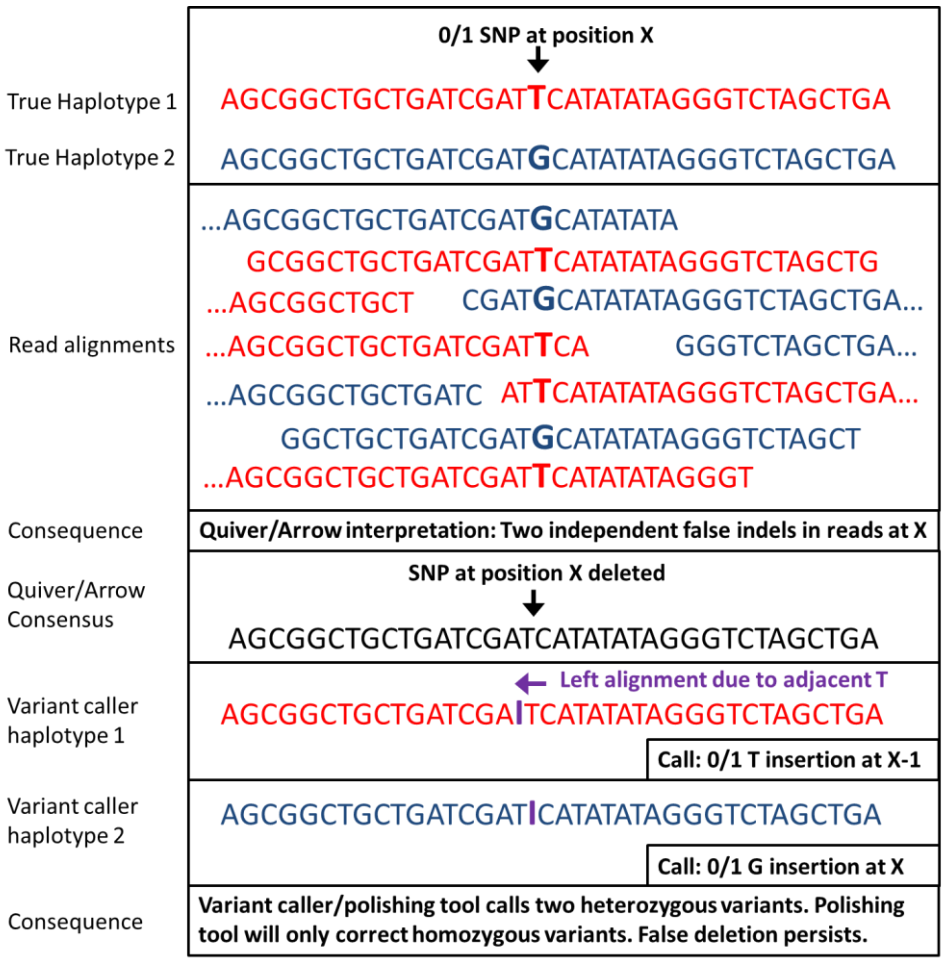



**Figure 5-1- Figure describing an issue with PacBio-only error correction with Quiver and Arrow on unphased genomes. Heterozygous SNPs are interpreted as two false indels in the raw data that occur at the same locus. The variant is deleted from the consensus sequence and in certain sequence contexts this appears to complicate further polishing efforts.**

### 5.2.9 Additional corrections to the assembly

Following identification of false indels introduced during polishing, these were corrected using indel-apply (https://github.com/douglasgscofield/PacBio-utilities Accessed 17/04/2019). Additionally Pilon was run 2 more times as described previously to correct any other errors that may have been missed the first time. Following this correction the Illumina data was aligned once more to assess how many homozygous variants remain. Additionally cDNA locations were downloaded from BioMart (Smedley et al., 2015) and the number of homozygous variants that overlap these were identified. Identified variants were examined to identify any reasons these had not been corrected previously.

## 5.3 Results

### 5.3.1 Initial read statistics

The read statistics following PacBio sequencing are detailed in table 5-1.

### 5.3.2 Initial assembly statistics

After the raw data had been assembled with Falcon, the assembly statistics were as described in table 5-2.

### 5.3.3 Contig quality control and contig splitting

A method adapted from the one described in chapter 2 was used to assess the quality of the contigs, this included the mapping of Illumina data, BAC ends and fosmids. Regions where there was a high amount of abnormal mapping and there was consistent evidence of problematic regions across the technologies were visualised in IGV and split where appropriate.

**Table 5-1- Statistics for raw reads following PacBio sequencing**

| | |
|---|---|
| **Number of reads** | 12,328,735 |
| **Total length of reads (bp)** | 175,934,815,397 |
| **Mean read length (bp)** | 14,270 |
| **Read N50 (bp)** | 19,786 |
| **Coverage** | 65X |

**Table 5-2- Statistics for initial assembly of Falcon contigs**

| | |
|---|---|
| **Number of haploid contigs** | 3,206 |
| **Contig N50 (Mb)** | 14.5 |
| **Largest contig (Mb)** | 63 |

**Table 5-3- Table showing results of RepeatMasker analysis on Sscrofa11.1**

| | | Number of Elements | Length occupied (bp) | Proportion of genome (%) |
|---|---|---|---|---|
| **SINEs** | | 1768010 | 354332615 | 14.16 |
| | **Alu/B1** | 0 | 0 | 0 |
| | **MIRs** | 410028 | 59957948 | 2.4 |
| **LINEs** | | 955168 | 521004536 | 20.82 |
| | **LINE1** | 644366 | 441431373 | 17.64 |
| | **LINE2** | 264322 | 69436494 | 2.78 |
| | **L3/CR1** | 34866 | 7480012 | 0.3 |
| | **RTE** | 10536 | 2488991 | 0.1 |
| **LTR elements** | | 306107 | 116921282 | 4.67 |
| | **ERVL** | 78412 | 33089318 | 1.32 |
| | **ERVL-MaLRs** | 129843 | 43715780 | 1.75 |
| | **ERV_classI** | 74712 | 32656740 | 1.31 |
| | **ERV_classII** | 4940 | 3004330 | 0.12 |
| **DNA elements** | | 291574 | 59974216 | 2.4 |
| | **hAT-Charlie** | 168861 | 32242758 | 1.29 |
| | **TcMar-Tigger** | 46031 | 12807393 | 0.51 |
| **Unclassified** | | 4376 | 806294 | 0.03 |
| **Total interspersed repeats** | | | 1053038943 | 42.09 |
| **Small RNA** | | 1359041 | 294287877 | 11.76 |
| **Satellites** | | 3317 | 5866795 | 0.23 |
| **Simple repeats** | | 755370 | 30567934 | 1.22 |
| **Low complexity** | | 123236 | 5873390 | 0.23 |
| **Total masked** | | | 1095772350 | 43.8 |

This was only the case for 28 of the 3,206 contigs with the majority appearing to be good quality or only have abnormal mapping in one technology (generally Illumina alignments in repetitive sequence).

### 5.3.4 Scaffolding and gap filling

346 contigs covering 2.3 Gb were included in the initial chromosomal scaffolds which contained 410 gaps. Following gap filling 108 gaps remained in 125 placed contigs.

All five BACs sequenced on Nanopore's MinION assembled into a single circular contig with Canu. All of these mapped to the expected positions and after removing the BAC vector sequences the insert sequences were placed to close the targeted gaps, leaving 103 gaps and 122 placed contigs in the final Sscrofa11 assembly and closing acrocentric chromosomes 16 and 18.

### 5.3.5 RepeatMasker

The results of the RepeatMasker analysis are listed in table 5-3.

### 5.3.6 Assembly quality control

The same Illumina dataset used in chapter 2 was aligned to Sscrofa10.2 and Sscrofa11.1 using BWA mem (v0.7.15). Table 5-4 shows the mapping statistics for the two assemblies according to Samtools (v1.3.1) flagstat. The assembly was assessed using the methods described in chapter 2 and the results are summarised in table 5-5.

**Table 5-4- Mapping rates for the same dataset between Sscrofa10.2 and Sscrofa11.1**

| | Sscrofa10.2 | Sscrofa11.1 |
|---|---|---|
| **Percent mapped** | 77.05% | 97.4% |
| **Percent properly paired** | 73.06% | 83.75% |

**Table 5-5- Results of LQLC analysis described in chapter 2 for Sscrofa10.2 and Sscrofa11**

| | % genome (Sscrofa10.2) | % genome (Sscrofa11) |
|---|---|---|
| **High Coverage** | 2.6 | 4.9 |
| **Low Coverage** | 26.6 | 7.5 |
| **Low % Properly paired** | 4.95 | 3.9 |
| **High % large inserts** | 1.52 | 1.72 |
| **High % small inserts** | 3.99 | 4.7 |
| **LQ** | 13.85 | 11.6 |
| **LQLC** | 33.07 | 16.3 |

**Table 5-6- Table showing the homozygous variants called by GATK between Illumina data from the reference animal and the reference genome**

| | Indel | SNP | Mixed | Total |
|---|---|---|---|---|
| **GATK** | 202,593 | 38,701 | 10,657 | 251,951 |
| **GATK less repeat regions** | 128,859 | 29,943 | 6,730 | 165,532 |

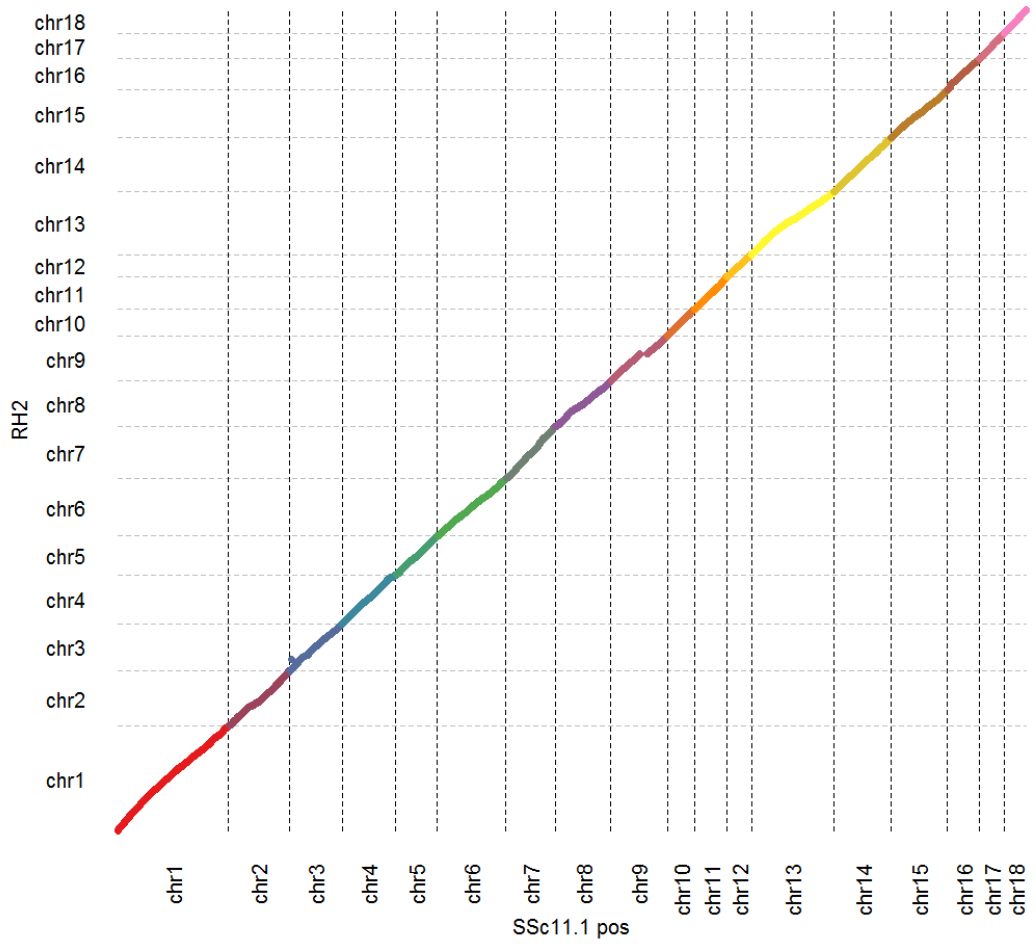



**Figure 5-2- Figure showing alignment positions of markers against Sscrofa11.1 compared to the expected position defined by an RH map.**

Several categories of LQ (LQ includes high coverage, low % properly paired, high % with large inserts and high % with small inserts) have increased compared to the percentage in Sscrofa10.2, however the total LQ has reduced suggesting these categories are more likely to overlap in Sscrofa11.1, likely due to the increase in incorporation of repetitive content. When the regions which have been flagged as LQLC are compared with regions identified by RepeatMasker (v4.0.6; Smit, 2013-2015), only 1.6% of the genome is LQLC and not in a repetitive region.

Figure 5-2 shows an alignment between Sscrofa11.1 and a radiation hybrid map (Servin et al., 2012). The two datasets are collinear with only a small discrepancy visible on chromosome 3 and an apparent disagreement on chromosome 9. The discrepancy on chromosome 3 is currently unexplained and likely represents a small misassembly, but the gap on chromosome 9 represents a gap in the RH map as described by Servin et al. (2012).

The analysis of SNPs and indels between the Sscrofa11 assembly and the reference animal focussed on homozygous variants, as heterozygous variants likely represent true variants in the reference animal and contain one allele which supports the reference. Homozygous variants are sequences where all the Illumina data support a particular base or insertion/deletion, but the assembly supports a different base or insertion/deletion state at the relevant location.

Table 5-6 shows a summary of the results from GATK including the raw homozygous variant calls and those which do not intersect repeat regions

defined by RepeatMasker. Accepting 165,532 as the number of individual base errors in the assembly gives an error rate of 0.007%. This error rate does not include the additional errors discovered after annotation and described below.

### 5.3.7 Final assembly statistics

Ideograms showing the difference in contiguity between Sscrofa10.2 and Sscrofa11 are shown in figure 5-3.The assembly statistics of Sscrofa10.2, Sscrofa11 and Sscrofa11.1 (Sscrofa11 with added Y chromosome) are detailed in table 5-7.

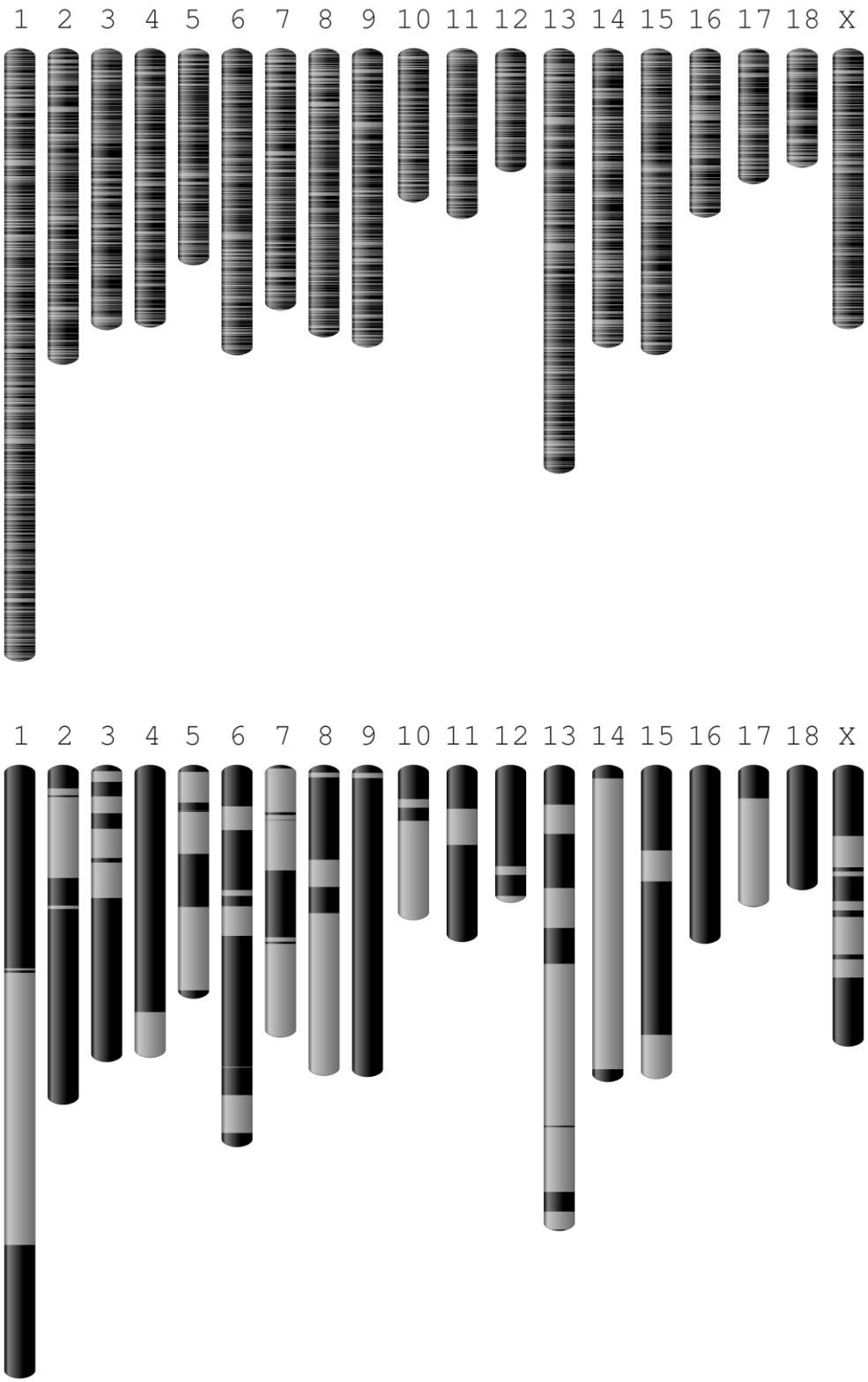



**Figure 5-3- Ideograms for Sscrofa10.2 (top) and Sscrofa11 (bottom) produced using NCBI genome decoration page (https://www.ncbi.nlm.nih.gov/genome/tools/gdp Accessed 19/06/18). For each ideogram, the change in colour indicates a switch of contig.**

### 5.3.8 External assessments of quality

This section describes work not done by the author of this thesis which assesses the quality of the work described above. Responsible parties are credited in text.

#### *5.3.8.1 gEVAL*

The gEVAL pipeline run by William Chow and Kerstin Howe at Wellcome Sanger Institute has a measure of contiguity called an assembly badge consisting of three numbers, XYZ, where $10^X$ represents the contig N50 in kb, $10^Y$ represents the scaffold N50 in kb, and Z is either 1 or 0 for assembled to chromosome level or not, respectively. The assembly badge for Sscrofa10.2 was 121 and the assembly badge for Sscrofa11.1 is 441, matching the assembly badge for mouse (GRCm38) and human (GRCh38.p5). Additionally, BUSCO (v2; Simão et al., 2015) was used to estimate the completeness of the assembly. The results for this using 4,104 mammalian BUSCOs with comparisons to other assemblies can be seen in table 5-8.

The gEVAL pipeline also allows for visualisations of alignments of BAC and fosmid end sequences and some cDNA sequences to the genome and comparisons between genomes. These visualisations will be discussed further later to demonstrate corrections between the assemblies.

The full gEVAL results and browser for this assembly can be accessed at http://geval.sanger.ac.uk/Pig_Ss11_1/Info/Index?db=core (Accessed 04/06/2018 ).

**Table 5-7- Final assembly statistics for Sscrofa10.2, Sscrofa11 and Sscrofa11.1 N50- more than 50% of the assembled sequence is in a contig/scaffold of X bases or greater. L50- half of the assembled sequence is in the largest X contigs/scaffolds.**

| | Sscrofa10.2 | Sscrofa11 | Sscrofa11.1 |
|---|---|---|---|
| **Total sequence length** | 2,808,525,991 | 2,456,768,445 | 2,501,912,388 |
| **Total assembly gap length** | 289,373,899 | 1,869,354 | 29,864,641 |
| **Gaps between scaffolds** | 5,323 | 24 | 93 |
| **Number of scaffolds** | 9,906 | 626 | 706 |
| **Scaffold N50** | 576,008 | 88,231,837 | 88,231,837 |
| **Scaffold L50** | 1,303 | 9 | 9 |
| **Number of contigs** | 243,021 | 705 | 1,118 |
| **Contig N50** | 69,503 | 48,231,277 | 48,231,277 |
| **Contig L50** | 8,632 | 15 | 15 |
| **Total number of chromosomes** | 21 | 19 | 21 |
| **Number of component sequences (WGS or clone)** | 186,661 | 705 | 1,308 |

**Table 5-8- BUSCO results from the gEVAL pipeline for Sscrofa10.2, Sscrofa11.1, the mouse reference genome (GRCm38), and the human reference genome (GRCh38.p5).**

| | Sscrofa10.2 | Sscrofa11.1 | GRCm38 | GRCh38.p5 |
|---|---|---|---|---|
| **Complete** | 80.9% | 93.8% | 95.2% | 94.9% |
| **Complete and single copy** | 80.2% | 93.3% | 91.6% | 94.1% |
| **Complete and duplicated** | 0.7% | 0.5% | 3.6% | 0.8% |
| **Fragmented** | 8.2% | 3.5% | 2.3% | 2.5% |
| **Missing** | 10.9% | 2.7% | 2.5% | 2.6% |

### *5.3.8.2 Cogent*

Cogent analysis uses PacBio Iso-seq data to assemble the transcriptome and identify gene families, which are then aligned to the genome assembly. If these assembled genes do not align to the assembly they are classified as missing from the assembly. This work was carried out by Elizabeth Tseng from PacBio and results are currently unpublished, but will be published in the manuscript for the work described in this chapter. This assessment was carried out using high-quality isoform sequences from 7 tissues (diaphragm, hypothalamus, liver, skeletal muscle (*longissimus dorsi*), small intestine, spleen and thymus). Cogent partitioned 276,196 high quality isoform sequences into 30,628 gene families, of which 18,708 have at least 2 sequences. Cogent focusses analysis entirely on genes with multiple isoforms. Cogent then performed reconstruction on the 18,708 partitions. For each partition, Cogent reconstructs transcribed ‘contigs’ that represent the ordered concatenation of transcribed exons as supported by the isoform sequences. The reconstructed contigs were then mapped back to Sscrofa11.1 and contigs that could not be mapped or mapped to more than one position were individually examined. The analysis found 5 genes missing: CHAMP1, ERLIN1, IL1RN, MB, and PSD4. Of these 5 genes, only ERLIN1 was present in Sscrofa10.2. Investigation of these missing genes was carried out by the author of this thesis. Two of the missing genes, IL1RN and PSD4, were present in the original set of PacBio contigs from Falcon, however they were trimmed during quality control. CHAMP1 is known to occur in a telomeric region on chromosome 11 and is likely erroneously

missing from the assembly. The genes expected to surround ERLIN1 are present on chromosome 14, however one of these neighbouring genes, CYP2C33, is duplicated and this may represent a misassembly. In the location where MB is expected to be located, there is no gap in the assembly, but only 14% of MB is present (according to the cDNA alignment in gEVAL). This region of the genome on chromosome 5 originated from a single Falcon contig and likely represents a misassembly by Falcon.

### 5.3.9 Annotation

Results of the annotation by Ensembl with comparisons to Sscrofa10.2 are presented in table 5-9. The assembly and annotation are available in Ensembl's genome browser (http://www.ensembl.org/Sus_scrofa/Info/Index, accessed: 06/06/18).

The results of the annotation by NCBI are presented in table 5-10 with comparison to Sscrofa10.2 and the annotated genome is available in NCBI's Genome Data Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?acc=GCF_000003025.6&context=genome, accessed 06/06/18) .

**Table 5-9- Annotation results from Ensembl for Sscrofa10.2 and Sscrofa11.1**

| | Sscrofa10.2 | Sscrofa11.1 |
|---|---|---|
| | **Ensembl (Release 89)** | **Ensembl (Release 92)** |
| **Coding genes** | 21,630 (Incl. 10 read through) | 22,452 |
| **Non-coding genes** | 3,124 | 3,250 |
| **small non-coding genes** | 2,804 | 2,503 |
| **long non-coding genes** | 135 (incl 1 read through) | 361 |
| **misc. non-coding genes** | 185 | 386 |
| **Pseudogenes** | 568 | 178 |
| **Gene transcripts** | 30,585 | 49,448 |
| **Genscan gene predictions** | 52,372 | 46,573 |
| **Short variants** | 60,389,665 | 64,310,125 |

**Table 5-10- NCBI annotation results for Sscrofa10.2 and Sscrofa11.1**

| Feature | Sscrofa10.2 | Sscrofa11.1 |
|---|---|---|
| **Genes** | 36,396 | 27,250 |
| **All transcripts** | 66,378 | 78,200 |
| **mRNA** | 47,251 | 63,562 |
| **misc_RNA** | 1,931 | 4,438 |
| **miRNA** | 425 | 395 |
| **tRNA** | 479 | 510 |
| **lncRNA** | 16,289 | 9,292 |
| **rRNA** | 2 | 2 |
| **SRP_RNA** | - | 1 |
| **Single-exon transcripts** | 2,225 | 2,409 |
| **coding transcripts (NM_/XM_ )** | 2,223 | 2,406 |
| **non-coding transcripts (NR_/XR_ )** | 2 | 3 |
| **CDSs** | 47,251 | 63,562 |
| **Exons** | 275,591 | 283,679 |
| **in coding transcripts (NM_/XM_ )** | 232,840 | 253,270 |
| **in non-coding transcripts (NR_/XR_ )** | 49,326 | 50,421 |
| **Introns** | 232,796 | 240,989 |
| **in coding transcripts (NM_/XM_ )** | 203,210 | 219,801 |
| **in non-coding transcripts (NR_/XR_ )** | 35,809 | 40,936 |
| **Mean transcripts per gene** | 1.82 | 2.92 |

### 5.3.10 Detection and correction of errors post annotation

An example of a heterozygous SNP that has been deleted during polishing is shown in figure 5-4. The figure shows an A/G SNP where the top and bottom track are the heterozygous sire and dam of the homozygous (G/G) offspring in the middle track. The deleted A allele appears as an insertion which has been left aligned due to the adjacent A in the reference, with the G appearing as an insertion in the correct position for this SNP. The reference animal is heterozygous at this locus.

Through Platypus variant calling, 471,064 pairs of these false heterozygous indels were identified including 3,433 in the coding regions affecting 1,763 genes. These were corrected using indel-apply from PacBio-utilities (https://github.com/douglasgscofield/PacBio-utilities; Accessed 17/04/2019).

The additional iterations of Pilon reduced the homozygous SNPs and indels further as described in table 5-11.

Of the remaining homozygous indels, 850 of them are in the coding region affecting 757 genes. The reason these have not been corrected was investigated and it was found that the majority of remaining variants have very low coverage in the Illumina data (figure 5-5). Most of these would likely be corrected with higher coverage Illumina data, but some are expected to remain in repetitive regions. These will be addressed by obtaining higher coverage Illumina data and correcting with Pilon.

**Figure 5-4- An IGV visualisation showing a heterozygous SNP for which the reference base has been deleted shown as two adjacent indels (purple). The top and bottom track show heterozygous individuals (A/G) and the middle track shows a homozygous individual (G/G).**

**Table 5-11- Table showing reduction of homozygous variants (likely reference base errors) following repeated runs of Pilon following NCBI submission**

| | Homozygous SNPs | Homozygous indels |
|---|---|---|
| **Sscrofa11.1** | 38,701 | 202,593 |
| **Additional Pilon run 1** | 37,152 | 63,844 |
| **Additional Pilon run 2** | 33,175 | 60,548 |

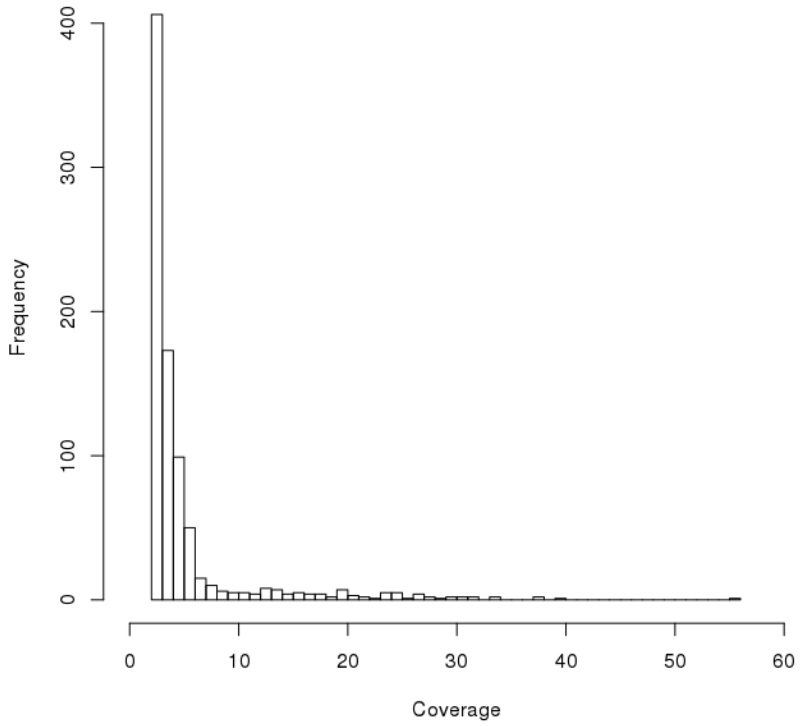



**Figure 5-5- Histogram showing the number of remaining homozygous variants at each coverage level in coding regions.**

### 5.3.11 Comparisons with Sscrofa10.2

As both Sscrofa10.2 and Sscrofa11.1 have been analysed using the gEVAL pipeline, it is possible to use this tool to visualise the corrections of loci identified as potentially misassembled in previous chapters and other researchers' work. Additionally the correction of fragmented genes and the inclusion of missing genes can be visualised to see how these are improved in Sscrofa11.1.

A region discussed in chapter 2.2 on chromosome 6 that was very clearly misassembled in Sscrofa10.2 is now corrected in Sscrofa11.1. The rearrangement of the misassembled region between the two genome assemblies is shown in figure 5-6. Several neighbouring genes have also been corrected in this region. Additionally, BAC end alignments now support the new arrangement of this region.

Genes that were misassembled, missing or fragmented beyond usefulness can also be seen corrected in the new assembly, two examples are shown in figure 5-7.

Figure 5-8 shows the region in Sscrofa10.2 that was identified as wrongly placed on chromosome 15 instead of chromosome 3 during the GWAS in chapter 3.3.4. This places the significant variant that was at 15:138464069 in Sscrofa10.2 in a region where a number of variants were approaching significance for number stillborn on chromosome 3.

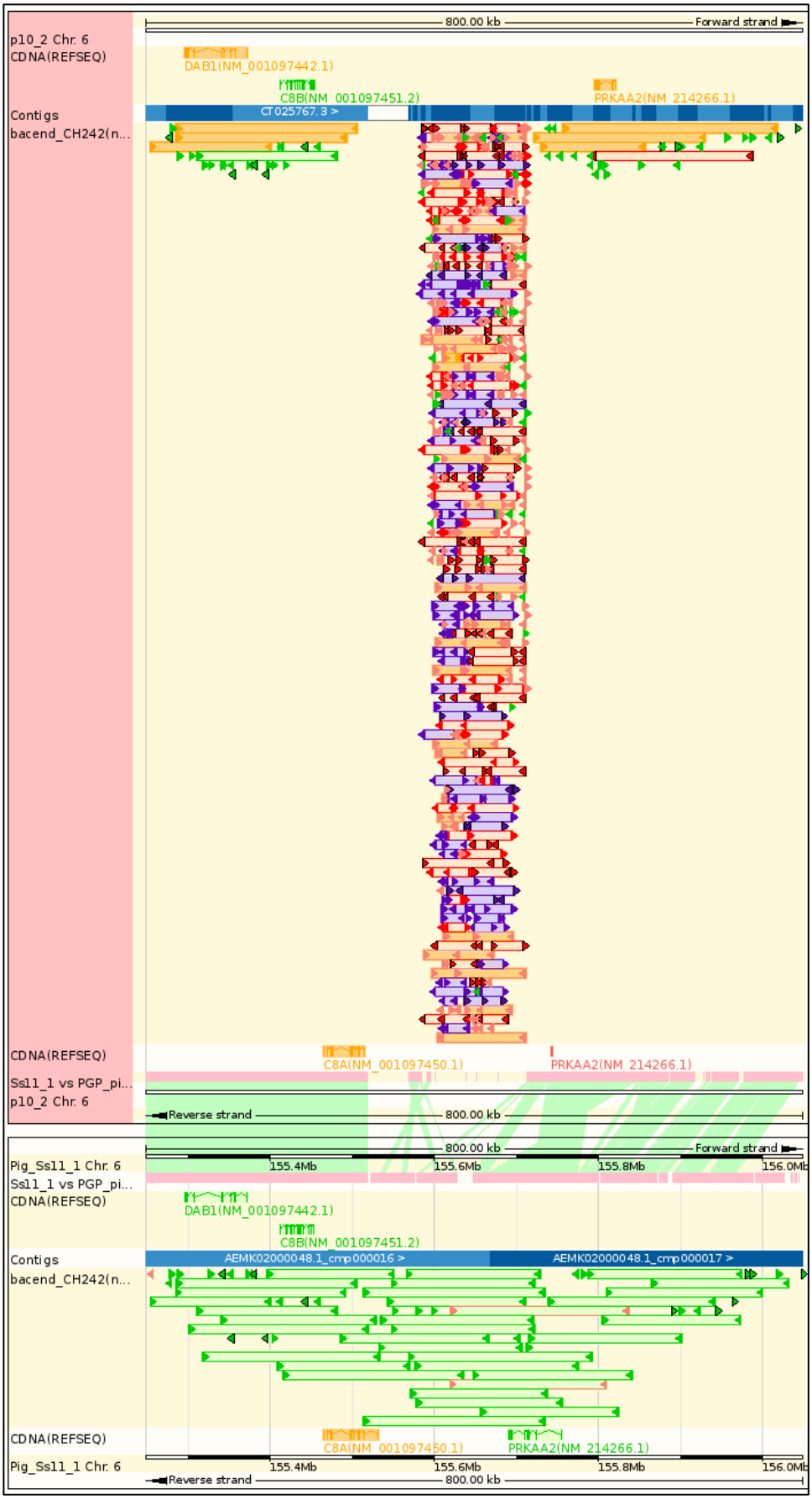
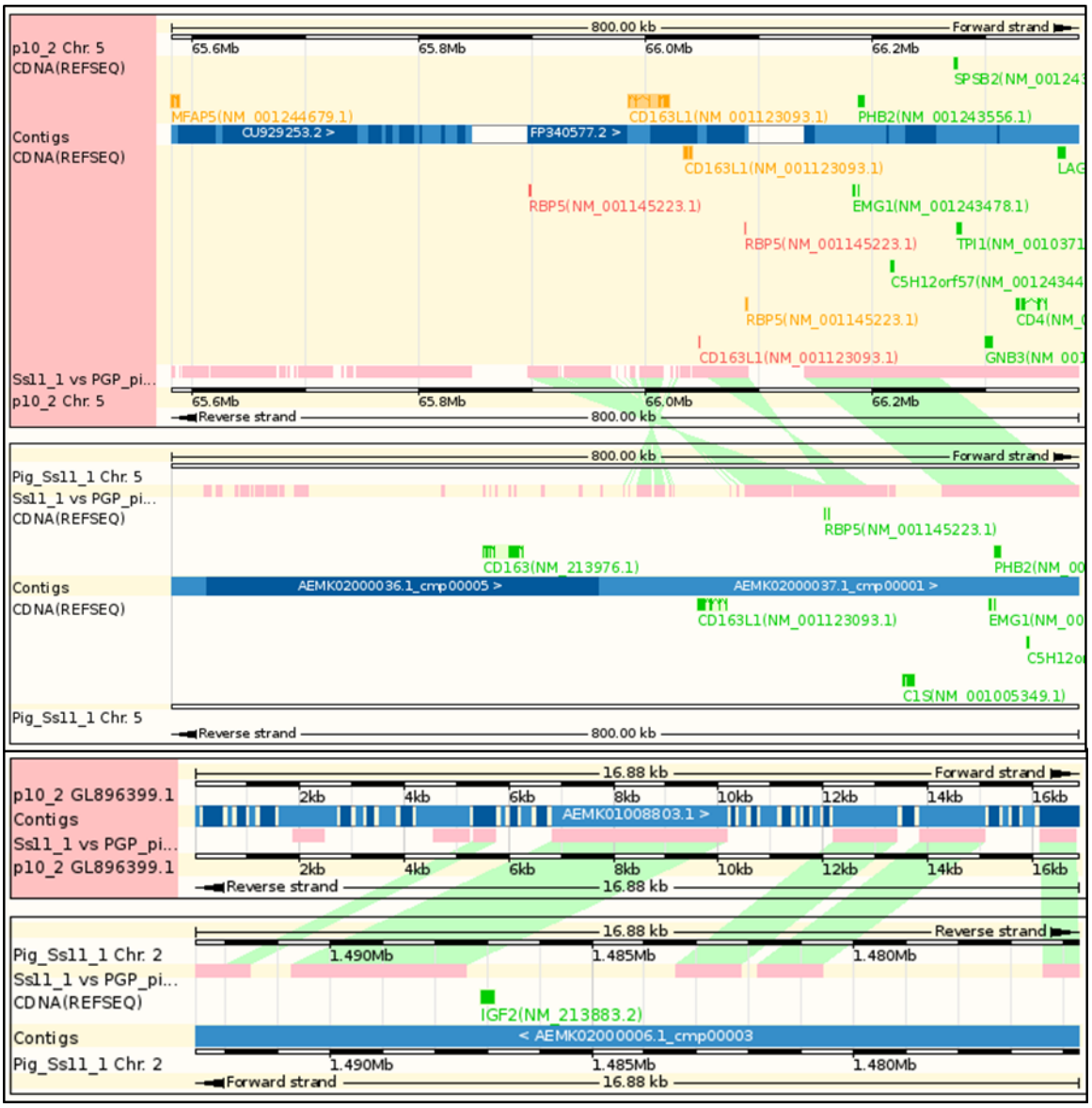



**Figure 5-6- gEVAL visualisation showing a misassembled region in Sscrofa10.2 (top), and its corrected region in Sscrofa11.1 (bottom). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies. Horizontal bars represent isogenic BAC end data aligned to the two assemblies, for which green alignments are accurate, orange have incorrect insert sizes, red are wrongly oriented, and purple are multimapped ends. Similarly, the CDNA track shows cDNA alignments, with the green genes having high coverage and identity.**

A




B




**Figure 5-7- gEVAL visualisation showing misassembled regions in Sscrofa10.2 (top track) around genes CD163 (A) and IGF2(B), and their corrected regions in Sscrofa11.1 (bottom track). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies. The CDNA track shows cDNA alignments, with the green genes having high coverage and identity. CD163 is likely missing due to the gap in Sscrofa10.2, while IGF2 is in a highly fragmented unplaced scaffold.**

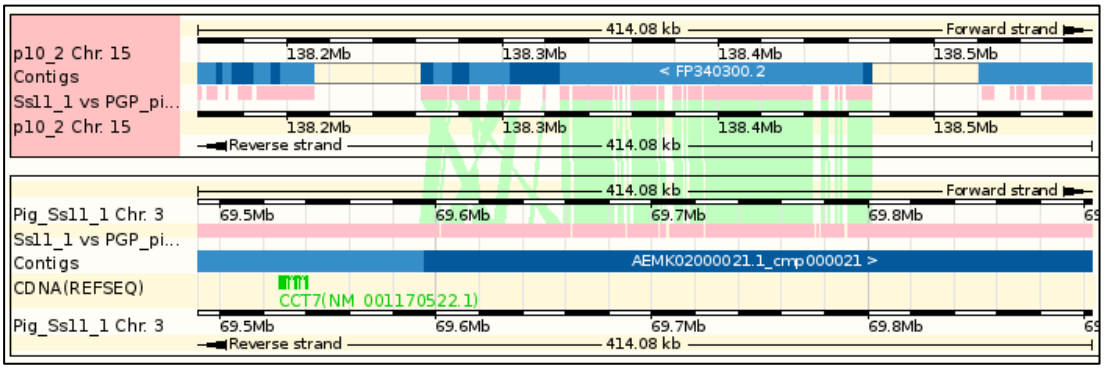



**Figure 5-8- gEVAL visualisation showing a misplaced contig in Sscrofa10.2 (top track) and its corrected position in Sscrofa11.1 (bottom track). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies. This contig on chromosome 15 in Sscrofa10.2 should be on chromosome 3.**

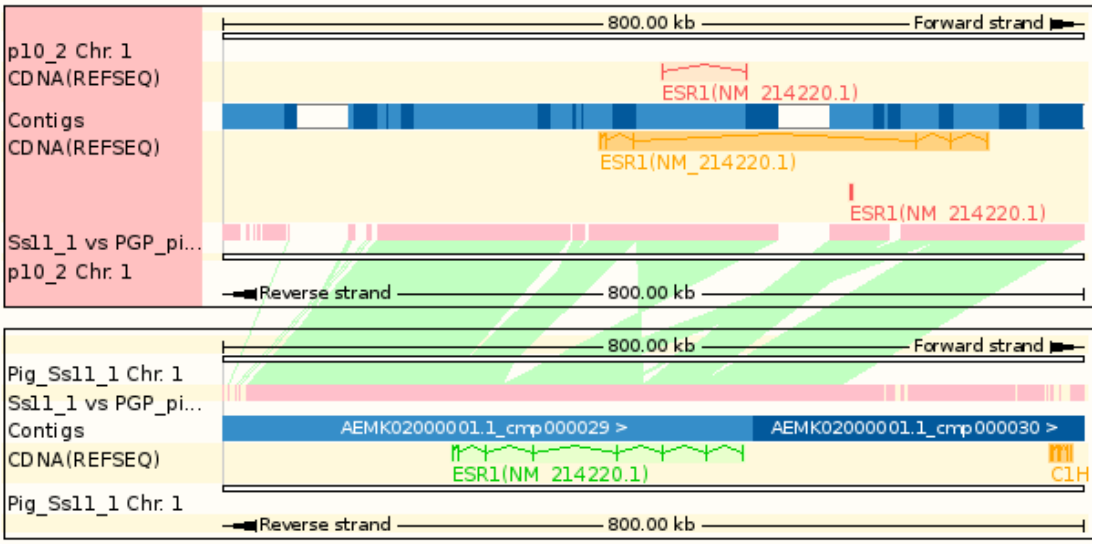



**Figure 5-9- gEVAL visualisation showing the misassembly of ESR1 in Sscrofa10.2 (top track) and its corrected structure in Sscrofa11.1 (bottom track). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies. Note that some exons of ESR1 are on the wrong strand in Sscrofa10.2.**

A region associated with litter size identified by this thesis in chapter 3.3.4 both supported and opposed by the results of others (Munoz et al., 2007, Rothschild et al., 1996, Short et al., 1997, Dall'Olio et al., 2011, Ma et al., 2012) containing the ESR1 gene, has been rearranged (figure 5-9) which may contribute to finding the causative variant of this debated QTL.

A region containing a variant of interest discussed in chapter 4.4 in which the variant was located in fragmented gene ENSSSCG00000025104, thought to be a part of ATP5H, is now correctly assembled and the missing portion of the gene was likely lost in an adjacent gap (figure 5-10). Additionally, neighbouring gene MRPL58 missing in Sscrofa10.2 has also been placed correctly.

Finally, a region on chromosome 4 associated with fatty acid composition suspected of being on the wrong chromosome by both van Son et al. (2017) and Yang et al. (2013) was indeed misplaced and is now in the region they predicted it should be located near SCD on chromosome 14 (figure 5-11) neighbouring another QTL for fatty acid composition.

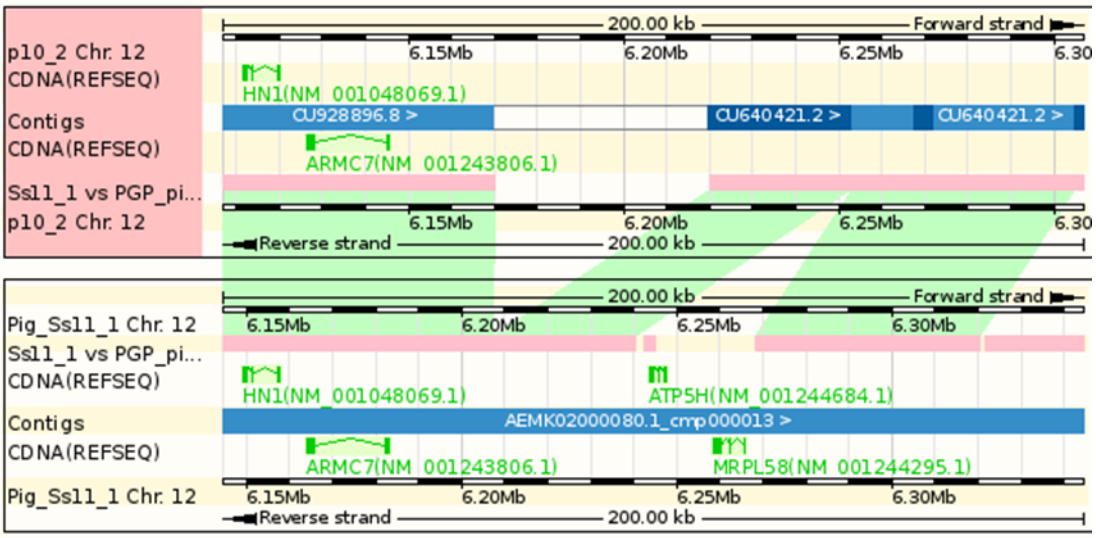



**Figure 5-10- gEVAL visualisation showing the region ATP5H is missing from in Sscrofa10.2 (top track) and its placement in Sscrofa11.1 (bottom track). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies.**

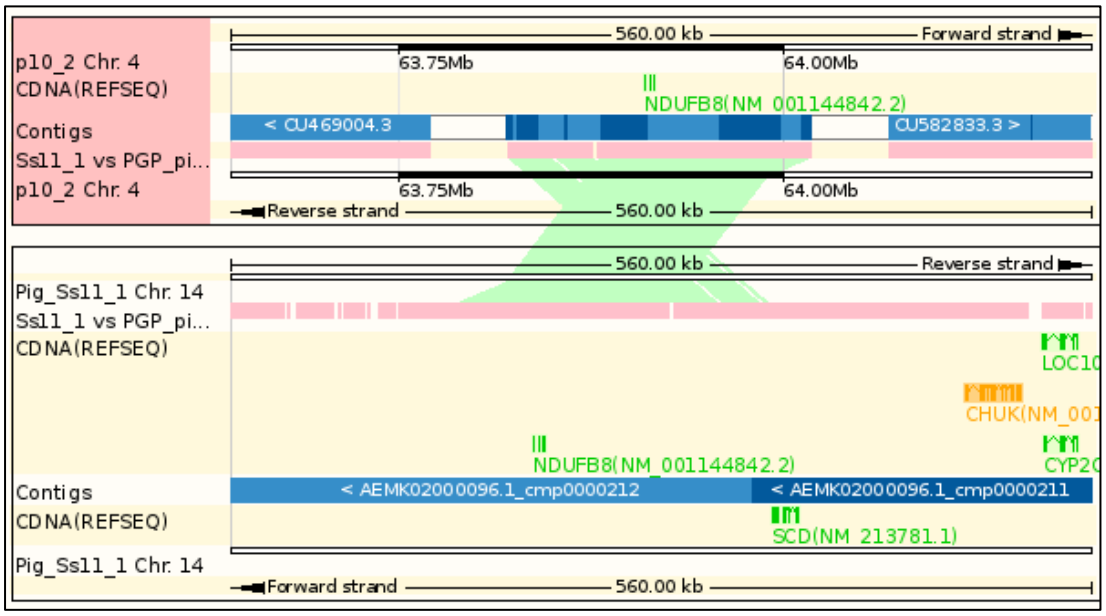



**Figure 5-11 gEVAL visualisation showing a misplaced contig in Sscrofa10.2 (top track) and its corrected position in Sscrofa11.1 (bottom track). Green lines between the two tracks show the alignment and rearrangement of sequence between the two assemblies. This contig on chromosome 4 in Sscrofa10.2 should be on chromosome 14.**

## 5.4 Discussion

Long read sequencing greatly improves our ability to assemble large genomes. With long-read technologies that are capable of sequencing individual reads longer than the contig N50 of earlier generations of genome assemblies, the challenge of assembling up to chromosome arm level is greatly reduced. For Sscrofa11.1, the assembly greatly benefited from a wealth of data already produced for the assembly of Sscrofa10.2. Many recent genome assemblies lack data such as physical maps to assist with assigning contigs to chromosomes. New technologies such as Hi-C, Dovetail, and optical mapping (Bionano) can be used to scaffold genomes and assign them to chromosomes. In this case, such additional data did not need to be produced thanks to the previous efforts to assemble the pig genome which were largely held back by the limitations of sequencing technologies of the time.

The new pig genome assembly presented here, Sscrofa11.1, is a substantial improvement over Sscrofa10.2. Perhaps the most obvious improvement is in the contig N50 of 48.2Mb, which has almost a 700-fold improvement over Sscrofa10.2, greatly reducing the potential for duplicated or fragmented gene models as demonstrated by the reduction in pseudogenes detected by Ensembl between Sscrofa10.2 and Sscrofa11.1. The contig N50 is also substantially higher than the goat genome with an N50 of 26.2Mb, which is considered the best livestock genome to date (Bickhart et al., 2017). Indeed, the contig N50 is comparable to that of the current human reference genome sequence (GRCh38.p12), on which considerably more work has been carried

out, with a contig N50 of 56.4Mb. When considering only the contigs assembled from long reads (Sscrofa11 rather than Sscrofa11.1), there are fewer gaps than the corresponding human genome chromosomes. Also of note are two closed chromosomes, 16 and 18, these chromosomes are acrocentric and are in a single contiguous sequence with no gaps. Contigs on chromosomes 4 and 9 appear to span their respective centromeres, however the centromeric sequence in the assembly at these loci probably represent a collapsed version of these highly repetitive sequences. The improved mapping rate of the same dataset between Sscrofa10.2 and Sscrofa11.1 may be both due to the increased contiguity and the inclusion of sequence missing from the previous assembly.

External assessments of the quality of Sscrofa11.1, including gEVAL, Cogent, and the two annotations suggest the genome is highly complete. While the BUSCO alignments provided by the gEVAL pipeline suggest there are still up to ~6% missing or fragmented genes, the BUSCO publication notes that in large genomes these two categories may be inflated (Simão et al., 2015), and these figures are much improved from Sscrofa10.2 and comparable to those reported for human and mouse. The Cogent analysis from PacBio reported just 5 genes missing, several of which could be traced back to the point where they were lost in the assembly. The majority of these missing genes originated from repetitive locations and had been misassembled by Falcon, or erroneously removed during contig quality control. The two annotations differ due to the different datasets and pipelines used to produce them, however both suggest fewer pseudogenes and

fragmented genes, representation of more isoforms and better representation of the coding sequence in general. The non-coding genes are still underrepresented in the annotation of this new pig genome when compared to human and mouse. However, this omission is likely attributable to the conservative nature of the Ensembl annotation system.

The improvement in the accuracy of the assembly will have a positive impact on analyses. For regions previously placed on the wrong chromosomes, the corrections in Sscrofa11.1 will simplify association analyses, where regions may have previously been searched for candidate genes that were in fact not associated with the trait in question (Yang et al., 2013, van Son et al., 2017). While a couple of examples have been described here that were identified as misassembled during analysis, it is likely that there are other examples that were not previously noticed. Additionally certain genes that may be of particular interest to the field, such as CD163 and ESR1, are now correctly represented in the genome. CD163 is an important gene for Porcine Reproductive and Respiratory Syndrome susceptibility (Calvert et al., 2007), a disease with a huge economic impact on the pig breeding industry. This locus has been used successfully as a target for genome editing to reduce susceptibility (Burkard et al., 2017, Burkard et al., 2018). ESR1 has been a subject of debate since 1996, with multiple papers either supporting or opposing its association with litter size in pigs (Munoz et al., 2007, Rothschild et al., 1996, Short et al., 1997, Dall'Olio et al., 2011, Ma et al., 2012). An accurate representation of this gene and the surrounding region in the reference may allow for a causative variant to be identified.

One of the major flaws with long read sequencing technology is the high error rate (Watson, 2018). Despite employing multiple polishing techniques, Sscrofa11.1 still has more indels than an assembly produced with a more accurate sequencing technology would. Some of the remaining indels were introduced by the polishing tools Quiver and Arrow: a known issue with bases at heterozygous loci being deleted. Pilon likely calls these deletions as two indels in much the same way that other variant callers such as GATK and Platypus do, and so does not correct them as it only corrects high confidence homozygous variants. However, some of the remaining indels have not been corrected due to low coverage in the Illumina data used to polish the assembly with Pilon. These errors will cause problems for calling SNPs and indels, particularly those in the coding region. While variants that are called in all individuals in a study are likely to be disregarded, indels in the assembly may cause truncation of gene models. If gene models are truncated, variant annotation will also be affected and benign variants may be annotated as impactful and detrimental variants that occur after a truncation in the gene model will be missed. Efforts to improve these issues for a future release have been described here and should greatly reduce these problems. However, some genes in low coverage regions are currently still affected.

Despite its flaws, long read sequencing offers an exciting opportunity to assemble highly contiguous reference genome assemblies. As prices continue to fall, accuracy improves and obtainable read lengths increase, even the smallest labs should be able to afford to assemble a high quality reference genome for their species of interest. In the future, perhaps whole

chromosomes will be sequenced in a single read, negating the need for genome assembly.

### 5.4.1 Future work

In order to further improve the pig reference genome assembly, we intend to obtain higher coverage Illumina data for the same individual and further polish the assembly for a future release. Additionally, plans are in place to further improve contiguity. There are only 103 gaps in Sscrofa11, and just 24 of those are not spanned by a BAC from the CHORI-242 library. By sequencing the BACs that span gaps, and BACs that reach into the unspanned gaps, we can fill or extend into these using similar methods to those in section 5.2.2.5. Additionally, another less contiguous PacBio pig genome assembly, USMARCv1.0 (Accession: GCA_002844635.1) has recently been released. This can be used to identify potential structural misassemblies such as the one Cogent identified at the MB locus. These may represent true structural variants, or suggest a misassembly. USMARCv1.0 may be used to identify targets that can be examined further using other data from T.J.Tabasco such as using the BAC library. Any regions suspected of misassembly can then be broken and a gap filler used to try to correct them, or can be sequenced across from BACs.

Following these corrections, the work carried out in chapter 3 and 4 of this thesis can be redone and should produce more reliable results. Additionally, many of the regions of the genome were excluded from these analyses either directly through filtering, or indirectly through inability to accurately annotate variants in poorly assembled regions. This means there are relatively

unexplored regions of the pig genome that may be associated with some of the traits of interest discussed in this thesis.

In the future, graph genomes are likely to become more widely used. Sscrofa11.1 is a highly contiguous assembly which can be used as a backbone for a graph genome, overlaying novel haplotypes. However, currently the community still does not have the tools and file formats needed to implement this.

### 5.4.2 Conclusions

With the increased use of genomic data in livestock breeding, the accuracy of available genomic resources becomes increasingly important. In chapter 2 of this thesis, the problems associated with the current pig reference genome were explored. In subsequent chapters, the difficulties of working with an inaccurate reference genome were highlighted. In this chapter a new, more accurate reference genome has been produced. While the assembly is not perfect due to lingering false indels and a small number of structural errors, it is a large improvement on the previous assembly and its contiguity is comparable to some of the best currently available large reference genomes.

# CHAPTER 6: DISCUSSION

---

*“There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.”*
-J.R.R. Tolkien, The Hobbit

Analyses to identify genes with an impact on traits of interest are traditionally focussed on identifying genomic regions associated with those traits. Generally this involves identifying a population or pedigree where quantifiable variation can be seen in the trait of interest, generating detailed phenotypic data for the individuals in the population, and genotyping the population for a large number of genetic polymorphisms. Once these datasets have been obtained, they can be tested for associations between the genotypes and the trait of interest. This allows for identification of regions of the genome that are associated with the trait and markers in the region that can be used to predict the trait of an individual. This can be used in a number of applications including diagnosis of disease and marker assisted selection in farmed animals and plants. These analyses benefit greatly from a draft reference genome assembly to assist with identifying haplotypes and regions of interest.

The reducing cost of genome sequencing, however, allows for large cohorts of individuals to be sequenced at the exome- or genome-wide level, allowing for new approaches to be taken to identify causal variants that affect trait variation and better understand how these traits are influenced by the variant(s). Through sequencing of large cohorts it has been revealed that protein-truncating, or “loss of function”, variants are fairly common, with each healthy human harbouring roughly 100 protein-truncating variants with around 20 in the homozygous form (MacArthur et al., 2012, MacArthur and Tyler-Smith, 2010). However, the majority of these will likely occur in non-essential or redundant genes. Those protein-truncating variants in essential

genes will most likely be seen mostly in the heterozygous form, but when carried in the homozygous form may be deleterious and potentially embryonic lethal except in cases where there is redundancy in gene function.

The original aim of this postgraduate research project was to identify putative loss of function variants in the pig genome, and to determine whether or not these may underlie embryonic loss in the pig. Through the use of multiple technologies including exome sequencing, imputation, GWAS, and whole genome sequencing, several candidates were identified. However, the analyses suffered from being carried out against a low-quality, draft reference genome. To improve the future potential of such analyses, the latest in long-read sequencing technologies was used to assemble a new, highly contiguous reference genome.

This work has identified a number of putative causative variants in genes with links to reproductive phenotypes and embryo mortality, including variants that have been identified using more than one method such as a variant in EXOG found in the 96 exomes in chapter 3 and the mummified foetus trio in chapter 4, and a variant in SERPINA3 that was selected as a candidate through bioinformatic filtering and is in a region associated with number stillborn by GWAS in chapter 3. Additionally, candidate genes associated with reproductive phenotypes were identified, however, the small sample size available for association analyses means that these require further validation. For a few of the candidate genes, past studies have identified the same gene or loci as candidates for the same phenotype, number stillborn or similar in pigs, these may be particularly interesting to follow up on using a targeted

approach to identify a causative variant. These include an association between number stillborn or live litter size and the oestrogen receptor 1 locus (ESR1) in previous studies (Munoz et al., 2007, Rothschild et al., 1996, Short et al., 1997, Alfonso, 2005), a region containing GABA receptors associated with number stillborn in a previous study (Schneider et al., 2015), a region containing ARHGAP24 is close to a previously identified QTL for prenatal survival (Hernandez et al., 2014), and a region containing MYOM2 overlaps a previously identified QTL for number stillborn (Onteru et al., 2012).

The candidate variants identified in this work could be further investigated by carrying out experimental crosses of individuals carrying the variants and genotyping the resulting litters to identify deviations from the expected Mendelian inheritance pattern. Alternatively, if heterozygous carriers are identified, data may already exist in the form of EBVs that could suggest differences between their reproductive success and that of wild-type individuals in the same commercial line.

All of the methods of identifying candidates described here would be more successful in larger cohorts. Larger cohorts increase the likelihood of observing rare, but non-lethal variants in the homozygous form, increase the accuracy of imputation and the power of GWAS, and improve the filtering potential when looking for a variant causative of a specific phenotype. The costs of whole genome sequencing have decreased greatly in recent years, however at current costs it is still not feasible to sequence sufficiently large cohorts in this way for most species. Exome sequencing is still a valuable tool for identifying variants in the protein-coding region in a cohort, while this

comes with the disadvantage of not sequencing the regulatory regions of the genome, our ability to understand the impact of variants in these regions is still limited and it is likely that even with whole genome sequencing data, analyses of this sort would first focus on the coding region. Additionally the methods used in this work would likely be more successful through targeting individuals from pedigrees that are suspected of carrying traits that are demonstrably heritable.

Researchers using genome sequencing data frequently rely on publicly available resources to analyse their data. While the availability of these resources has been of benefit to the community, it is important that the limitations of these resources is known and accounted for. In this thesis I have described the extent of the problems of publicly available reference genome Sscrofa10.2, with its many structural errors including sections of sequence placed on the wrong chromosome. These problems have a large impact other important resources with over half of dbSNP variants for this genome being called from low-confidence regions. Additionally, the impact of misassemblies in Sscrofa10.2 on analyses can be seen in multiple genomic methods including wrongly placed GWAS SNPs obscuring associations in chapter 3, and false positives in exome sequencing and whole-genome sequencing in chapters 3 and 4. Sscrofa10.2 also suffers from missing sequence, including missing genes. While some of these have been analysed using targeted approaches, many of these will represent unexplored regions of the pig genome and may be relevant to important phenotypes. Additionally, missing sequence has consequences for methods

such as genome editing, where unique targets must be designed for the region to be modified, if sequence missing from the reference genome is highly similar to a target sequence this could increase incidence of unwanted off-target effects.

This work has provided the community with a new reference genome with far fewer problems than the previous assembly, with very few gaps remaining and excellent contiguity there is far less missing sequence. Many genes that were known to be missing are now in their expected genomic positions, and reads and markers from SNP chips can be more accurately mapped which will likely improve the accuracy of imputation and improve ease of interpretation of GWAS. While the Sscrofa11.1 assembly is an improvement on the previous one, some errors still remain as a consequence of the sequencing technology and polishing methods used. Further work will need to be done to polish the assembly to increase its reliability for variant calling. Additionally, BACs have been identified that span many of the remaining gaps and sequencing of these is under way to further increase the contiguity of an already highly contiguous assembly. With these final error corrections and gap closures the genome will be an extremely valuable resource to researchers working on pig genetics and genomics. Following these final corrections and reannotation of the genome, the exome and whole genome sequencing analyses described here can be repeated and should contain fewer false-positive calls and enable the analysis to cover regions of the genome that were previously inaccessible.

As sequencing technologies continue to improve in accuracy and read length it may be possible to further improve the genome assembly by sequencing across particularly repetitive regions such as the centromeres, as has recently been demonstrated using ONT sequencing to assemble the human Y chromosome centromere (Jain et al., 2018b), and produce a completely closed linear genome.

*De novo* assembly, even with long reads, is a computationally and monetarily expensive process. Reference-guided assembly, assuming availability of a high-quality reference, is an attractive approach to assembling the genomes of different breeds and closely related species with lower cost and fewer resources required. Chinese pig breeds are known to have genomes that differ from European breeds, and previously a draft genome assembly from short reads had been produced to assist with analyses on Chinese pigs (Fang et al., 2012). The availability of Sscrofa11.1 may facilitate the assembly of a better reference genome for Chinese breeds through reference-guided assembly. Researchers working on other species such as mice, cattle and dogs have begun to assemble reference genomes for multiple breeds, the use of which may more accurately represent certain populations, this too is an option for the pig to represent more breeds than simply separate European and Chinese assemblies. In the future, graph genomes may be more widely used to represent species, incorporating these different breeds and in-breed variation, rather than assembling these *de novo*, a high-quality linear reference genome such as the one presented here can be used as a backbone.

Long-read sequencing has greatly reduced the challenge of genome assembly, and as read lengths increase allowing long repetitive regions to be spanned by individual reads, closed linear assemblies will be achievable. The remaining problems with long reads centre around their high error rate compared to short-read sequencing technologies. Polishing currently is still best done using short reads, which will likely remove true variation in repetitive regions or between subtly different genes, and may fail to polish regions with a high number of multimapper or otherwise low-mappability regions at all. If polishing can be done solely with long reads this would help to reduce this problem as the reads would be anchored by unique sequence either side of these regions. If the problem of a high error rate can be addressed it will further simplify the assembly process by reducing the need for polishing. Currently both PacBio and Nanopore are working on solving this problem. PacBio are focussing on improving their PacBio-only polishing tool, Arrow, to reduce the error rate in homopolymers and those caused by diploid error (Drake, 2018). ONT are working on improving their base calling tools to reduce errors at an earlier stage, and additionally have plans to improve consensus accuracy through the use of multiple pore types with different error profiles (Brown, 2018). Some of the errors in ONT may relate to base modifications affecting the shape of the squiggle produced during sequencing, and if this is the case the base caller needs only to be trained to recognise these. However, similarly to PacBio the error rate in homopolymers is still high, it is hoped that through pores with longer constriction points, or

multiple constriction points, the base caller will be able to more accurately estimate the length of the homopolymer.

Genome editing in pigs offers exciting opportunities for the improvement of pork production. Once variants causative of traits have been identified and an accurate reference assembly is available for the design of guide RNA to target regions, genome editing can be used to greatly accelerate beneficial changes in the genome. Recently, pigs that have had a small edit made to the gene CD163 have been shown to be resistant to the Porcine Reproductive and Respiratory Syndrome virus, an extremely costly disease challenging the global pork production industry. The pigs appear otherwise to be normal, healthy pigs (Burkard et al., 2018). CD163 is one of the genes missing from Sscrofa10.2. Similarly successful edits have been made in other livestock species, such as hornless, or ‘polled’, dairy cattle (Carlson et al., 2016). Currently, however, it is unclear whether food products from genome edited individuals or their offspring will be allowed to be sold (Jones, 2015, Voytas and Gao, 2014, Ruan et al., 2017), nor is it clear if the public would accept such products (Ishii and Araki, 2016) given the continued strong resistance against genetically modified organisms despite scientific evidence that these are safe (Blancke et al., 2015, Marques et al., 2014). Another application of gene editing in pigs is for medical purposes. This may involve replicating variants that are causative of human diseases to make medical models, or editing to make pigs safe for use in xenotransplantation of cells, tissues or organs into humans. The similarity of pigs and humans make this technically possible, however a major barrier in its use is the presence of

porcine endogenous retroviruses (PERVs) in the pig genome. These PERVs are inactive in the pig and do not harm the host, but may become active and pathogenic in a recipient. Genome editing has been proposed as a method to inactivate the PERVs, and has been carried out on porcine kidney epithelial cells to disrupt 62 PERV genes and reduce transmission to human cells >1000-fold (Yang et al., 2015). Following this, genome editing has been used to knock out PERVs in pigs, revealing that pigs without PERVs are viable (Niu et al., 2017). However, PERVs are not the only barrier to xenotransplantation, and further edits may be needed to modulate the immune response and reduce risk of rejection (For review, see Hryhorowicz et al., 2017)

Pigs are an economically important species and the improvement of pig production efficiency is an important goal towards securing the global supply of pork. With an improved reference genome and the reduced cost of DNA sequencing, genomics offers powerful methods to understand the underlying causes of phenotypes important to the pig production industry.

# APPENDIX

*“That was it. That was really it.*
*She knew that she had told herself that*
*that was it only seconds earlier,*
*but this was now the final real ultimate it.”*
-Douglas Adams, Dirk Gently's Holistic Detective Agency

G3: Genes|Genomes|Genetics Early Online, published on July 2, 2015 as doi:10.1534/g3.115.018564



# Exome Sequencing: current and future perspectives

**Amanda Warr*, Christelle Robert*, David Hume*, Alan Archibald*, Nader Deeb§ and Mick Watson***
* The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK
§ Genus plc., Hendersonville, TN 37075, USA

**ABSTRACT** The falling cost of DNA sequencing has made the technology affordable to many research groups, enabling researchers to link genomic variants to observed phenotypes in a range of species. This review focusses on whole exome sequencing and its applications in humans and other species. The exome has traditionally been defined to consist of only the protein coding portion of the genome; a region where mutations are likely to affect protein structure and function. There are several commercial kits available for exome sequencing in a number of species and, owing to the highly conserved nature of exons, many of these can be applied to other closely related species. The data set produced from exome sequencing is many times smaller than that of whole genome sequencing, making it more easily manageable and the analysis less complex. Exome sequencing for disease gene discovery in humans is well established and has been used successfully to identify mutations that are causative of complex and rare diseases. Exome sequencing has also been used in a number of domesticated and companion species. The successful application of exome sequencing to crops has yielded results that may be used in selective breeding to improve production in these species, and there is potential for exome sequencing to provide similar advances in livestock species that have not yet been realised.

**KEY WORDS**
exome
genome
sequencing

## INTRODUCTION

The completion of a reference genome sequence for humans took more than 200 scientists over a decade in a project that cost almost $3 billion to complete (International Human Genome Sequencing Consortium 2004). Over the past decade the price of genome sequencing has plummeted, much of the work has become automated and methods have improved. These advances mean complete genomes can be sequenced quickly and affordably. Human genome sequencing is a special case and enjoys much lower cost-per-gigabase than other species due to Illumina's HiSeq X platform (Watson 2014). Sequencing can allow for the identification of genetic variants that affect heritable phenotypes including important disease-causing mutations and natural variation that can be exploited to improve crops and livestock. Despite the significant improvement in sequencing technology, sequencing whole genomes to a depth sufficient to find variants that affect phenotypic expression is expensive when compared with targeted sequencing. This review will focus on exome sequencing: a method which targets only a subset of the genome, often the protein coding portion, significantly reducing the sequencing space and subsequently the cost. Details of the method, available platforms, uses in humans and other species and its benefits over whole genome sequencing (WGS) will be discussed.


doi: 10.1534/g3.115.018564

Manuscript received Apr. 15, 2015; Accepted for publication May 22, 2015
[1]Corresponding author: Mick Watson, The Roslin Institute and Royal (Dick), School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK, Tel: +44131 6519208, Email: mick.watson@roslin.ed.ac.uk

## WHAT IS EXOME SEQUENCING?

The exome has been traditionally defined as the sequence encompassing all exons of protein coding genes in the genome and covers between 1% and 2% of the genome, depending on species. It may also be extended to target functional non-protein coding elements (e.g. miRNA, lincRNA etc.) as well as specific candidate loci. There are two main categories of exome capture technology: solution-based and array-based.

In solution-based whole exome sequencing (WES), DNA samples are fragmented and biotinylated oligonucleotide probes (baits) are used to selectively hybridise to target regions in the genome. Magnetic streptavidin beads are used to bind to the biotinylated probes, the non-targeted portion of the genome is washed away and the polymerase chain reaction (PCR) is used to amplify the sample, enriching the sample for DNA from the target region. The sample is then sequenced before proceeding to bioinformatic analysis. Array-based methods are similar except that the probes are bound to a high-density microarray. The array-based method was the first to be used in exome capture (Albert et al. 2007), but it has largely been supplanted by solution-based methods, which require less input DNA and are consequently potentially more efficient; however, studies by Asan et al. (2011) and Bodi et al. (2013) found that NimbleGen's Sequence Capture Array performed better than the solution-based alternatives in low GC content regions; had high sensitivity and read mapping rates; and single nucleotide polymorphism (SNP) detection from these reads was more specific to the target region. This suggests that a niche may remain for the older tech-

nology. Array-based capture has been used successfully and accurately to identify rare and common variants and identify candidate genes for monogenic diseases in small cohorts (Ng et al. 2009); however, the array-based methods are less scalable owing to the limitation of the number of probes that can be accommodated on the array and additional equipment and time required to process the microarrays.

## EXOME CAPTURE PLATFORMS

There are several differences between the available platforms, which are constantly being updated and improved. The major providers of exome capture platforms are NimbleGen, Agilent and Illumina and each have different designs and strengths (Summarised in Table 1). The discussion of the characteristics of each platform will focus on the solution-based human kits; the performance of kits for other species has not been subjected to the same level of comparison.

NimbleGen's SeqCap EZ Exome Library has the highest bait density of any of the platforms, utilising short (55-105 bp), overlapping baits to cover the target region (Clark et al. 2011). This approach has been found to be an efficient method for enrichment with the least amount of sequencing needed to cover the target region and sensitively detect variants (Clark et al. 2011) and also has a high level of specificity showing fewer off-target reads than other platforms (Clark et al. 2011; Sulonen et al. 2011). Importantly, this bait design has been found to show greater genotype sensitivity and more uniformity of coverage in difficult to sequence regions, such as areas of high GC content, than the other platforms (Asan et al. 2011; Sulonen et al. 2011; Bodi et al. 2013).

Agilent's SureSelect Human All Exon Kit is the only platform to use RNA probes, with the other platforms opting for DNA probes. The baits used are longer than those used in NimbleGen's platform (114-126 bp) and the corresponding target sequences are adjacent to one another rather than overlapping (Clark et al. 2011). This design has been found to be good at identifying insertions and deletions (indels) as longer baits can tolerate larger mismatches (Clark et al. 2011; Bodi et al. 2013; Chilamakuri et al. 2014); it has been suggested that this may also reduce reference allele bias at heterozygous sites compared to other bait designs, however in practice the allele bias has been similar to other platforms (Asan et al. 2011; Hedges et al. 2011). The platform has been found to produce fewer duplicate reads than NimbleGen, but also fewer high quality reads (Sulonen et al. 2011). Bodi et al. (2013) found that Agilent had a higher alignment rate and fewer PCR duplicates than NimbleGen, but also had less uniform coverage.

Illumina's TruSeq Exome Enrichment Kit uses 95 bp probes that leave small gaps in the target region, with paired end reads extending outside the bait sequence during sequencing to fill the gap. This design has been found to have a high percentage of off-target enrichment (Clark et al. 2011) which reduces its target efficiency compared to the other platforms. This kit detects more SNVs in the untranslated regions (UTRs) than the other platforms (Clark et al. 2011), though comparisons of performance with NimbleGen and Agilent's "+UTR" kits have yet to be performed. Following filtering for duplicates, multiple mappers, improper pairs and off-target reads, Chilamakuri et al. (2014) found that this platform retained fewer reads (54.8%) than NimbleGen (66%) or Agilent (71.7%). At high read counts (>50M), this platform outperformed Agilent's SureSelect in downstream identification of indels (Clark et al. 2011).

Illumina's Nextera Rapid Capture Exome and Expanded Exome kits are similar to the TruSeq kit in their probe design. They differ from the other kits in that they use transposomes to fragment the genomic DNA, whereas the other platforms use ultrasonication. These kits have not been extensively compared to the other platforms with only Chilamakuri et al. (2014) having included Nextera in a comparison study. At the time of the study there was only one Nextera kit which was the Expanded Exome version: the kit with the larger target region of the two. The Expanded Exome kit shares a target region with the TruSeq kit which includes UTRs and miRNAs. Chilamakuri et al. (2014) found that the Nextera kit had increased coverage of high GC content areas due to altered bias in the transposome technology used during fragmentation, decreasing its overall uniformity; however recent changes to the protocol in the new versions may have improved this. They also found that out of all platforms tested, the Nextera platform retained the fewest

**Table 1 Summary of the differences between the solution-based exome sequencing platforms**

| | NimbleGen's SeqCap EZ Exome Library | Agilent's Sure Select Human All Exon Kit | Illumina's TruSeq Exome Enrichment Kit | Illumina's Nextera Rapid Capture Exome Kit |
|---|---|---|---|---|
| Probe Size[a] | 55–105 bp | 114–126 bp | 95 bp | 95 bp |
| Probe Type | DNA | RNA | DNA | DNA |
| Coverage Strategy | High-density, overlapping probes | Adjacent probes | Gaps between probes | Gaps between probes |
| Fragmentation Method | Ultrasonication | Ultrasonication | Ultrasonication | Transposomes |
| Target region size (Human)[b] | 64Mb | 50Mb | 62Mb | 62Mb |
| Reads remaining following filtering[c] | 66% | 71.7% | 54.8% | 40.1% |
| Major strengths | (i) High sensitivity and specificity<br>(ii) Most uniform coverage in difficult regions | (i) Better coverage of indels<br>(ii) High alignment rate<br>(iii) Fewer duplicate reads than other platforms | (i) Good coverage of UTRs and miRNAs | (i) Good coverage of UTRs and miRNAs |
| Major weaknesses | (i) More duplicate reads than Agilent<br>(ii) Lower alignment rate than Agilent | (i) Fewer high quality reads than NimbleGen | (i) High off-target enrichment | (i) High off-target enrichment<br>(ii) Coverage bias for high GC content areas reducing uniformity |
| Non-human supported | ✔ | ✔ | ✘ | ✘ |

[a] As described in Clark et al. (2011)- NimbleGen SeqCap EZ v2.0, Agilent SureSelect Human All Exon 50Mb, Illumina TruSeq Exome Enrichment
[b] For NimbleGen SeqCap EZ v3.0, Agilent SureSelect V5, and Illumina TruSeq and Illumina Nextera original versions
[c] Filtering for duplicates, multiple mappers, improper pairs and off-target reads, data from Chilamakuri et al. (2014)

reads after filtering for duplicates, multiple mappers, improper pairs and off-target reads, at 40.1%.

Chilamakuri et al. (2014) compared the target regions of the human kits from three providers (NimbleGen SeqCap EZ v3.0 – 64.1Mb; Agilent SureSelect V4 – 51.1Mb; and Illumina TruSeq and Illumina Nextera original version and protocol – 62.08 Mb). Despite the fact that all of the platforms are targeting the human exome, there is surprisingly little overlap between the three designs with just 26.2 Mb covered by all three target regions. NimbleGen's and Agilent's coverage are more similar to one another than either are to Illumina's; this is likely due to the large amount of UTRs covered in Illumina's target region. Illumina has 22.5Mb of targets unique to their platforms and 21.8Mb of these are UTRs. NimbleGen and Agilent have 16.1Mb and 7Mb of unique targets, respectively. It would be interesting to see how NimbleGen and Agilent's "+UTR" kits compare to Illumina in SNV calling, particularly in UTRs.

Currently, in addition to human kits, NimbleGen offer capture kits for maize, barley, wheat, soy, mouse and pig exomes and Agilent offer capture kits for mouse, cattle and zebrafish exomes. Both providers also offer the opportunity to design custom kits for other species. The kits for non-human species use similar bait designs and protocols to the providers' human kits. Both manufacturers offer a flexible design process allowing for modifications to improve coverage for specific regions and purposes.

## USES IN HUMANS

The first successful use of WES to diagnose and inform subsequent treatment in a human patient was in the identification of the causal variant of a rare form of Inflammatory Bowel Disease (IBD) in an infant (Worthey et al. 2011). In this case conventional diagnostics had failed to find an explanation for the patient's severe symptoms and doctors needed to understand the underlying cause of the symptoms before they could decide how to treat the child. A multidisciplinary team combined clinical phenotyping, exome sequencing, bioinformatics and functional studies, eventually finding the causative mutation which influenced the future treatment of the child. The analysis of the exome data was hampered by the relative lack of software designed for this purpose at the time and researchers had to manually inspect over 2000 variants. Through filtering and manual inspection the candidate pool was reduced to 70 genes exhibiting hemi- or homozygous variants. Of these only eight variants were novel and predicted to be damaging to protein function. Analysis of evolutionary conservation identified two variants that were located in highly conserved sequences and one of these had a high null genotype frequency. This left a single hemizygous, non-synonymous variant in the *X-linked inhibitor of apoptosis (XIAP)* gene. The child was diagnosed with X-linked lymphoproliferative disease 2, exhibiting a novel IBD-like manifestation caused by loss of tolerance to commensal organisms in the digestive system. This diagnosis allowed for effective treatment through allogeneic hematopoietic progenitor cell transplant. The unusual manifestation of the condition meant that the patient was unlikely to have been diagnosed without the exome sequence data.

Following the success of this initial diagnosis, exome sequencing has been used extensively to diagnose novel diseases and find novel causative mutations for known disease phenotypes. Exome sequencing is useful in human medicine for diagnosis of particularly difficult to diagnose patients, diagnosis of young patients who may not yet exhibit a full spectrum of symptoms (Iglesias et al. 2014), prenatal diagnosis (Iglesias et al. 2014; Xu et al. 2014) and early diagnosis of debilitating disease (Bras and Singleton 2011; Sassi et al. 2014). In addition to reaching a diagnosis, finding the causative mutation can allow for alteration of treatment, prevention of further invasive testing, accurate prognoses, and confirmed diagnoses which are essential for eligibility for benefits and access to clinical trials (Grossmann et al. 2011; Rabbani et al. 2012; Taneri et al. 2012; Iglesias et al. 2014) and in the future may allow for more targeted treatment.

Exome sequencing in human medicine benefits from the availability of large databases of known SNPs, known pathogenic variants and control genomes; during analysis variants found in these databases can generally be excluded when looking for novel variants, significantly reducing the variant pool. The Exome Aggregation Consortium (ExAC) has created a user-friendly database containing the exome sequences of over 60,000 unrelated individuals, which is freely available (Exome Aggregation Consortium (ExAC) 2014). The exomes have been analysed using a uniform bioinformatic pipeline and are from individuals with adult-onset diseases. This provides researchers with a large set of reference exomes which should be free from homozygous variants that cause childhood-onset Mendelian diseases. The database provides a wealth of information such as depth of coverage, genotype quality, allele frequency and variant consequences. The filtering capabilities and large number of exomes will simplify the process of prioritising variants when using exome sequencing as a diagnostic tool, particularly in children.

Where many unrelated, affected individuals are available, an 'overlap' strategy can be used to simplify analysis and search for common variants likely to affect gene function (Johansson et al. 2012). If there is a known inheritance pattern this can be used to search for specific genotype zygosity. Sequencing of multiple affected related individuals can also increase the power of the analysis (Johnson et al. 2010; Shi et al. 2011). In rare diseases, case-parent trios can be used to exclude non-pathogenic variants found in the parents (Smith et al. 2014). However, single patient sequencing may be sufficient to identify the causative mutation (Wang et al. 2011; Worthey et al. 2011; Xu et al. 2014). Analysis examining how conserved an amino acid sequence is through evolution and between genes in a family can help to increase the confidence of a mutation having a deleterious effect (Wang et al. 2011; Johansson et al. 2012; Sassi et al. 2014; Smith et al. 2014; Xu et al. 2014).

Examples of diseases for which exome sequencing has been used to detect a causative variant include Leber Congenital Amaurosis (Wang et al. 2011), Alzheimer's Disease (Sassi et al. 2014), Maturity-Onset Diabetes of the Young (Johansson et al. 2012), High Myopia (Shi et al. 2011), Autosomal Recessive Polycystic Kidney Disease (Xu et al. 2014), Amyotrophic Lateral Sclerosis (Johnson et al. 2010), immunodeficiency leading to infection with human herpes virus 8 causing Kaposi Sarcoma (Byun et al. 2010), Acromelic Frontonasal Dystois (Smith et al. 2014) and a number of cancer predisposition mutations (e.g. Yan et al. 2011; Greif et al. 2012; Snape et al. 2012; Kiiski et al. 2014; Cai et al. 2015).

The Deciphering Developmental Disorders (DDD) project aims to develop a scalable exome sequencing workflow to facilitate the translation of the method from the research environment to a clinical environment. So far over 1,000 children with undiagnosed developmental disorders and their parents have been sequenced with plans to increase this number to 12,000 patients. With diagnostic yields up to 31% in the children sequenced so far, this project demonstrates the future potential of the method in a clinical setting (Gecz and Corbett 2015; Wright et al. 2015; The Deciphering Developmental Disorders 2015).

## USES IN OTHER SPECIES

### Variants discovery for agricultural improvement

Not all protein-altering variants cause disease (MacArthur and Tyler-Smith 2010). To understand the potential of exome sequencing in areas other than disease variant discovery, which is the main focus in human research, we can look to studies using the technique to sequence agricultural species.

Plant genomes can be extremely complex, repetitive and are often polyploid; as a result some of the most economically important crops are not well suited for genome re-sequencing studied. For example, Bread Wheat (*Triticum aestivum*) has an allohexaploid (AABBDD) genome around 17Gb in size (Brenchley et al. 2012; The International Wheat Genome Sequencing Consortium 2014). This is likely too large for subsequent WGS studies at the current price of sequencing and data storage. Wheat is an extremely important crop for both human and livestock consumption and genetic improvement is slow; with the human population increasing and additional challenges from environmental changes it is essential to gain a better understanding of the crop in order to improve it. An exome capture kit has been designed for wheat based on the accumulated transcriptome data (Winfield et al. 2012). The capture region for this kit is 56.5Mb, which is around the lower estimated size of one diploid wheat exome, and owing to similarity between the three genomes may be sufficient to capture most of the exome data from the whole allohexaploid genome. This kit has been used to identify induced mutations in the genome to aid studies investigating gene function, a use that was also applied to the rice (*Oryza sativa*) exome in the same study (Henry et al. 2014) and the soybean (*Glycine max*) exome in a separate study (Bolon et al. 2011). WES in soybean has also been used to identify unwanted intracultivar genetic heterogeneity in the exome that may affect the plant's phenotype (Haun et al. 2011).

The genome of barley (*Hordeum vulgare* L.) has not been fully sequenced. Barley's genome is smaller than wheat's at around 5Gb, but is still larger than is practical for routine whole genome sequence analysis and contains repetitive elements that complicate the genome's assembly. A gene space assembly has been produced (The International Barley Genome Sequencing Consortium 2012) and a barley exome capture kit has been developed based on this assembly (Mascher et al. 2013). The kit has since been used to identify a mutation involved in early maturation, a trait relevant to production (Pankin et al. 2014).

Exome capture in barley has also been used to identify a gene causative of many-noded dwarfism (mnd) using mapping-by-sequencing (Mascher et al. 2014). The mnd phenotype is a shorter plant with more, narrower leaves than the wild type. The mutant in this study was created using X-ray mutagenesis; a technique which often causes large deletions. An F2 population between mutant and wild type phenotypes was created and 18 mutant individuals and 30 wild type individuals were exome sequenced. From these sequences SNPs were identified and allele frequencies of these were used to identify an allele over represented in the mutant group. Researchers queried the sequencing reads for exome targets that were present in the wild type but not the mutant. This lead to the identification of a candidate gene (MLOC_64838.2, now HvMND), which has a homolog known to play a role in a similar phenotype in rice. Screening of other mutants showing this phenotype found a variety of null mutations in this gene. The family to which this gene belongs is known to have effects on important production traits and may include good selection targets to improve production.

## Identification of new genetic markers

Robert et al. (2014) have designed exome capture probes for the pig, used these to sequence the exomes of 96 healthy pigs and identified potentially deleterious variants. Bioinformatic analysis identified 236,608 high confidence predicted variants and 28,115 predicted indels in the target region. This work revealed notable gaps in the current Ensembl *S.scrofa* genome annotation and identified a large number of potential protein truncating variants. As the pigs tested were healthy, it is possible that some of these protein truncating variants have phenotypic effects on traits other than those relating to the health of the pigs or to the production traits currently under selection. This work is an important step in identifying phenotype altering variants in the pig: a production animal and medical model.

The barley exome kit has been used to differentiate between markers of *H. vulgare* L. and *H. bulbosum* L.; *H. bulbosum* L. is a wild species that has superior pathogen resistance and tolerance compared to the domestic species and the two can be crossed to improve the domesticated crop, however negative linkage drag on production traits has hampered its use in elite barley lines. Using exome sequencing to identify specific markers can allow selective crossing to be used to incorporate the beneficial variants without incorporating linked variants that are detrimental to production (Wendler et al. 2014). Similarly, the wheat exome kit has allowed for discovery of previously unidentified markers in the genome which can be used in future genetic studies and marker assisted selection (Allen et al. 2013).

Black cottonwood (*Populus trichocarpa*) has had an exome capture kit designed (Zhou and Holliday 2012). *P. trichocarpa* is a model organism and was the first tree to have its whole genome sequenced (Tuskan et al. 2006). The tree is used in lumber production and in cosmetics. The tree has experienced a whole genome duplication fairly recently and the exome study found that this does not appear to have had an effect on SNP detection through exome sequencing (Zhou and Holliday 2012). There is potential to use the identified markers to improve production in this species.

## Health traits

As with exome sequencing in humans, exome sequencing has been used in other mammals to discover variants associated with health traits. WES has been used in conjunction with a genome wide association study to identify a frameshift mutation causing blindness in Phalène dogs (Ahonen et al. 2013). In cattle (*Bos taurus*), WES has been used successfully to identify strong candidate variants for haplotypes relating to reduced fertility rates in Holsteins which can be used to selectively breed against these detrimental haplotypes (McClure et al. 2014).

## Special considerations for non-human species

For many species, the reference genome is not completed to the same standard as the human genome; many species have only a draft genome or, as discussed for barley, no reference genome. This is an important consideration when using WES in non-human species. Poor annotation of genomes mean that in the design of capture probes, causative genes may be missed because they are not annotated whereas with whole genome sequencing, the data will be there whether the gene is annotated or not. Where available, including predicted genes identified from RNA-sequencing data may be beneficial to maximize coverage of functional elements in poorly annotated genomes. For example, Robert et al. (2014) added an additional 14Mb of data to the capture region in pigs using EST evidence. Additionally, errors in the reference genome will greatly increase the number of false-positive variant calls in these species. This increases computational burden and forces more stringent filtering which may inadvertently discard causative variants.

WES may be especially useful for model organisms, particularly where the sequences of large numbers of individuals are needed. The expense of WGS on large cohorts makes it less feasible in animal studies than human studies, particularly in species with fairly large genomes such as mice (~3.5Gb) and zeberafish (~1.5Gb), both common model organisms. Sequencing a smaller portion of the genome, particularly if candidate regions are known and can be targeted, would be more cost-effective, allow deeper coverage and potentially increase the number of individuals that can be used in these studies.

The genomes of animals have different levels of linkage disequilibrium (LD), which are often high in domestic species and model organisms with limited effective population sizes. The level of LD may also

vary by breed, as demonstrated in the domestic dog (Stern et al. 2013). When using exome sequencing to identify causal variants, high LD may lead to the identification of a benign variant that is in LD with the causative variant. In this case, the causative variant may have been missed in variant calling or lie outside the sequence space; it is therefore important to consider the known function of the element the candidate variant is associated with and the predicted effect of the variant on function. Additionally, efforts to sample individuals that are as outbred as possible may help to reduce this problem. Another consideration with high LD is that it may be possible to use this information in imputation, as has been done with other technologies such as imputing genotypes from SNP arrays (Hickey et al. 2012), from WGS data (Deelen et al. 2014; Gudbjartsson et al. 2015) and from human exome data (Auer et al. 2012). The larger haplotypes associated with high LD increase the accuracy and reduce the computational burden associated with imputation. Imputation may be less accurate than sequencing; however it allows for a larger numbers of individuals to be used at lower coverage and may help to reduce problems associated with variable coverage in exome sequencing.

## EXOME CAPTURE TRANSFERABILITY BETWEEN SPECIES

The Neanderthal exome was successfully sequenced by Burbano et al. (2010) using an exome capture kit designed for the human exome. The study compared the Neanderthal exome to human exomes and found 88 fixed substitutions in 83 genes in the human exomes; these substitutions did not appear to be a result of positive selection and may be a result of accelerated genetic drift from reduced effective population size in humans following historical bottlenecks or reduced purifying selection.

The kit designed for cattle can also be applied to other bovid species; Cosart et al. (2011) demonstrated that the kit could be successfully used to capture the exomes of, and identify SNPs in, zebu (Bos indicus) and American Bison (Bison bison). This transferability of exome capture kits, as also demonstrated in studies involving the sequencing of Neanderthals and non-human primates using human capture kits (Burbano et al. 2010; Vallender 2011), is possible because despite millions of years of divergence, functional elements tend to be highly conserved.

## WHY NOT USE GENOME SEQUENCING?

With the price of sequencing falling as rapidly as it has done over the past decade, questions have been raised concerning WES's usefulness in the era of affordable WGS. However, WGS is still more expensive than WES. The costs of WES consist of the cost of the capture plus the cost of sequencing, whereas WGS consists only of the sequencing costs. If we assume that the cost of capture remains fixed, then as the costs of sequencing fall, the cost of WGS will approach the cost of WES. However, at present that is not the case; and it would be unwise to assume that the cost of sequence capture will not reduce.

Human genomes are a special case, in that the HiSeq X platform offers a cost-per-Gb far less than other platforms, yet is limited (contractually, rather than technically) to 30X WGS human genomes (Watson 2014). However, even given that advantage, the $1000 price tag for a 30X human genome is (in our estimate) two- to three- times the cost of a 40X human exome (depending on scale). It may be advantageous to sequence more samples using WES, and gain statistical power, than to sequence more of the genome. In other species, the price difference is even higher – for example, in pigs (a similar sized genome to human but currently without the benefit of access to the HiSeq X platform) we estimate the cost of WGS to be 9-10 times the cost of WES (Robert et al. 2014).

While WGS does have benefits over WES, the cost of this technology is more than simply the price of sequencing. While sequencing technology has been improving at a much faster rate than would be predicted by Moore's law (a prediction of improvement in computing hardware, but often also applied to other technologies), the technology for storing and analysing the data has not seen a matching acceleration in improvement (Mardis 2010; Sboner et al. 2011). WGS produces around one hundred times the data that WES does at the same coverage. The infrastructure needed to store, manage and analyse data significantly increases the costs of WGS. WGS produces a much larger number of variants than WES does, not only because of the size of the sequencing space, but because regions outside the exome are less well conserved; while this number might include a variant of interest, the larger data set significantly increases the computational burden for analysis. Additionally variation in non-coding regions is less well understood than variation in the coding region (Mu et al. 2011; Maurano et al. 2012; Ward and Kellis 2012), making it more difficult to predict which variants might be relevant to a trait of interest in WGS datasets. The majority of causative variants identified so far in Mendelian disease have been found in coding regions (Botstein and Risch 2003), although ascertainment bias is likely to play a role in this conclusion. A study investigating functional non-coding variants based on WGS data from 1092 human genomes showed that functionally deleterious non-coding mutations were under strong negative selection, in a similar way to that of loss-of-function variants in protein-coding regions (Khurana et al. 2013). The authors developed a tool to prioritize non-coding variants in disease studies, which was used to identify non-coding candidate drivers in tumour genomes.

With the continuous decrease in sequencing cost, new studies making use of WGS to investigate causative variants will lead to the discovery of additional mutations in regulatory elements that contribute to the pool of disease-associated variants. In that context, the sampling bias currently observed towards coding variants is likely to be reduced by WGS investigations of non-coding genomic regions.

However, cost is not the only consideration. WGS covers the whole genome at more consistent coverage than WES, can provide more accurate detection of structural variants and does not have reference sequence bias caused by probe sequences in WES (Majewski et al. 2011; Meynert et al. 2014; Belkadi et al. 2015). Recent studies have highlighted and suggested roles for promoters (The Fantom Consortium et al. 2014) and enhancers (Andersson et al. 2014) in a range of different cell types, and these are not traditionally captured by exome sequencing. Importantly, WES requires prior knowledge of the location and sequence of features in order to target them, whereas WGS covers the entire genome. This means that exome sequencing relies on the accuracy of the genome annotation, so phenotype altering variants may be missed in poorly or incompletely annotated genomes. It is therefore important to take care when designing the capture region or purchasing a commercial kit to ensure any specific regions of interest are included. Exome sequencing also invariably fails to successfully capture the entire target region (Asan et al. 2011; Bodi et al. 2013; Chilamakuri et al. 2014; Robert et al. 2014), causing even properly annotated regions to be missed. Another consideration is that exome sequencing involves a PCR stage which is known to reduce coverage of GC-rich regions (Kozarewa et al. 2009; Veal et al. 2012). New sequencing technology has allowed for sequencing of DNA without the need for DNA amplification, generating sequencing from single molecules. This technology can produce accurate, longer reads without the artefacts and biases associated with the amplification process in other sequencing methods and the exome capture stage of WES (Shin et al. 2013). The longer read length in these third-generation sequencers is particularly beneficial for detecting structural variants and for resolving repetition in assemblies and copy number variable regions (Roberts et al. 2013). However, for now the price of this long-read sequencing is still prohibitively expensive and is not commonly in use for analysis of genetic variation. In the future as these sequencers improve and prices come down they may make WGS

more attractive to researchers even if their analysis focusses solely on the protein-coding region and known functional elements.

The power of WGS on a large cohort for variant discovery has already been demonstrated by Gudbjartsson et al. (2015), who sequenced the whole genomes of 2,636 Icelanders to 20X and imputed the sequence variants into 101,584 further chip-genotyped and phased individuals from the same population. This allowed for discovery of 6,795 null mutations in 4,924 genes which may play a role in disease. They also found evidence of lethal mutations that are not found in the homozygous state, which likely impact on fertility in heterozygous couples. Imputation reduces the amount of sequencing and data storage required, though the study found that fewer null mutations were detected in the imputed data set suggesting that some mutations were missed in these individuals.

WGS will eventually take a leading role in genome interrogation; however, it will likely have to wait for data storage and analysis to improve before its full potential can be realised. In the meantime, WES provides many of the benefits of WGS with lower storage requirements and computational burden, at an affordable price. This is particularly useful in large scale studies, for example, a recent study sequenced the exomes of over 9,000 people (Schick et al. 2015). It is also useful for the sharing of information such as in the ExAC database where there are over 60,000 exomes stored. WES will likely also remain the method of choice in species with exceptionally large genomes, for example in some polyploid plant species.

## CONCLUSIONS

Exome sequencing is a technology that allows interrogation of the most well understood portion of the genome: the protein-coding sequence and functional elements. Variants of interest don't necessarily fall within the exome, but so far most of the known variants responsible for Mendelian disease have been found in the coding region and the target region can be extended to include other regions of interest. While the falling price of sequencing may soon make genome sequencing more attractive, the additional costs of data handling and downstream analysis cannot be ignored. The applications of variant discovery, particularly in disease gene identification, cannot afford to wait for data storage and processing technology to catch up to sequencing improvements. The amount of data produced by WES is far more manageable than WGS; particularly for small research groups and groups studying organisms with large genomes. WES has established itself as an important method in disease gene identification in humans, and increasingly in domestic species. The applications of WES in crop research is allowing genomic techniques to be used in species with complex genomes, potentially identifying variants important for production that can be incorporated into marker-assisted selection. At present, WES is a useful and powerful method for variant discovery within coding regions offering most of the benefits of WGS while allowing for easier analysis and storage of the data produced.

However WGS will eventually be the NGS technology of choice with regard to the investigation of genomic variations due to the more uniform coverage achieved with WGS-versus WES- including coverage within the exome (Belkadi et al. 2015) and the more uniform distribution of sequencing quality parameters (*e.g.* coverage depth, genotype quality) observed with WGS- versus WES-(Meynert et al. 2014; Belkadi et al. 2015). Additionally WGS extends the variation search space to the whole genome allowing for the additional detection of non-coding variants (The 1000 Genomes Project Consortium 2012), the functional impact of which is becoming easier to interpret with the systematic annotation of functional non-coding elements (The ENCODE Project Consortium 2012; Ward and Kellis 2012; Andersson et al. 2014; The Fantom Consortium et al. 2014).

Ideally -regardless of sequencing costs and storage limitations-, integrated approaches combining the advantages of both WES and WGS would be beneficial for variant discovery studies with the addition of WES-/WGS-exclusive variants with WES providing additional variants missed in low-coverage dataset (The 1000 Genomes Project Consortium 2012). Additionally, both technologies' usefulness depend on the quality of the reference assembly that the sequenced reads are mapped to; poor quality references increase the number of false-positive variants identified in the analysis, which inevitably leads to more stringent filtering, increasing the potential for discarding a variant of interest. In the future, the improvement of reference assemblies, bioinformatic tools and sequencing technology will be necessary to improve the power of variant discovery techniques.

The term "exome" may no longer be appropriate for a technique which is simply a subset of the more generic technique of sequence capture. It is already possible to extend the exome capture region beyond protein coding genes to capture non-coding genes, and regulatory elements such as promoters and enhancers. In this way, we may sequence all of the functional areas of a genome without the cost of sequencing everything. Indeed, as clinically important variants are discovered, the paradigm may change to sequencing small panels of genes that are known to be relevant to disease, as is common in cancer genomics.

## REFERENCES


Ahonen, S. J., M. Arumilli, and H. Lohi, 2013 A CNGB1 frameshift mutation in Papillon and Phalene dogs with progressive retinal atrophy. PLoS ONE 8 (8):e72122.

Albert, T. J., M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler *et al.,* 2007 Direct selection of human genomic loci by microarray hybridization. Nat Meth 4 (11):903-905.

Allen, A. M., G. L. Barker, P. Wilkinson, A. Burridge, M. Winfield *et al.,* 2013 Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (Triticum aestivum L.). Plant Biotechnol J 11 (3):279-295.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt *et al.,* 2014 An atlas of active enhancers across human cell types and tissues. Nature 507 (7493):455-461.

Asan, Y. Xu, H. Jiang, C. Tyler-Smith, Y. Xue *et al.,* 2011 Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol 12 (9):R95.

Auer, Paul L., Jill M. Johnsen, Andrew D. Johnson, Benjamin A. Logsdon, Leslie A. Lange *et al.,* 2012 Imputation of Exome Sequence Variants into Population- Based Samples and Blood-Cell-Trait-Associated Loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet 91 (5):794-808.

Belkadi, A., A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent *et al.,* 2015 Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences.

Bodi, K., A. G. Perera, P. S. Adams, D. Bintzler, K. Dewar *et al.,* 2013 Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech 24 (2):73-86.

Bolon, Y. T., W. J. Haun, W. W. Xu, D. Grant, M. G. Stacey *et al.,* 2011 Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol 156 (1):240-253.

Botstein, D., and N. Risch, 2003 Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33:228-237.

Bras, J. M., and A. B. Singleton, 2011 Exome sequencing in Parkinson's disease. Clin Genet 80 (2):104-109.

Brenchley, R., M. Spannagl, M. Pfeifer, G. L. A. Barker, R. D/'Amore *et al.,* 2012 Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491 (7426):705-710.

Burbano, H. A., E. Hodges, R. E. Green, A. W. Briggs, J. Krause *et al.,* 2010 Targeted investigation of the Neandertal genome by array-based sequence capture. Science 328 (5979):723-725.

Byun, M., A. Abhyankar, V. Lelarge, S. Plancoulaine, A. Palanduz *et al.,* 2010 Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J Exp Med 207 (11):2307-2312.

Cai, J., L. Li, L. Ye, X. Jiang, L. Shen *et al.,* 2015 Exome sequencing reveals mutant genes with low penetrance involved in MEN2A-associated tumorigenesis. Endocr Relat Cancer 22 (1):23-33.

Chilamakuri, C. S. R., S. Lorenz, M.-A. Madoui, D. Vodák, J. Sun *et al.,* 2014 Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15:449-449.

Clark, M. J., R. Chen, H. Y. Lam, K. J. Karczewski, G. Euskirchen *et al.,* 2011 Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29 (10):908-914.

Cosart, T., A. Beja-Pereira, S. Chen, S. B. Ng, J. Shendure *et al.,* 2011 Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC Genomics 12:347.

Exome Aggregation Consortium (ExAC), 2014, Cambridge, MA. (URL: http://exac.broadinstitute.org) [Accessed: 04/2015]

Gecz, J., and M. Corbett, 2015 Developmental disorders: deciphering exomes on a grand scale. The Lancet 385 (9975):1266-1267.

Greif, P. A., A. Dufour, N. P. Konstandin, B. Ksienzyk, E. Zellmeier *et al.,* 2012 GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. Blood 120 (2):395-403.

Grossmann, V., E. Tiacci, A. B. Holmes, A. Kohlmann, M. P. Martelli *et al.,* 2011 Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. Blood 118 (23):6153-6163.

Gudbjartsson, D. F., H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson *et al.,* 2015 Large-scale whole-genome sequencing of the Icelandic population. Nat Genet advance online publication.

Haun, W. J., D. L. Hyten, W. W. Xu, D. J. Gerhardt, T. J. Albert *et al.,* 2011 The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155 (2):645-655.

Hedges, D. J., T. Guettouche, S. Yang, G. Bademci, A. Diaz *et al.,* 2011 Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform. PLoS ONE 6 (4):e18595.

Henry, I. M., U. Nagalakshmi, M. C. Lieberman, K. J. Ngo, K. V. Krasileva *et al.,* 2014 Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. Plant Cell 26 (4):1382-1397.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. J. van der Werf, and M. A. Cleveland, 2012 A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics, Selection, Evolution : GSE 44 (1):9-9.

Iglesias, A., K. Anyane-Yeboa, J. Wynn, A. Wilson, M. Truitt Cho *et al.,* 2014 The usefulness of whole-exome sequencing in routine clinical practice. Genet Med 16 (12):922-931.

International Human Genome Sequencing Consortium, 2004 Finishing the euchromatic sequence of the human genome. Nature 431 (7011):931-945.

Johansson, S., H. Irgens, K. K. Chudasama, J. Molnes, J. Aerts *et al.,* 2012 Exome sequencing and genetic testing for MODY. PLoS ONE 7 (5):e38050.

Johnson, J. O., J. Mandrioli, M. Benatar, Y. Abramzon, V. M. Van Deerlin *et al.,* 2010 Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron 68 (5):857-864.

Khurana, E., Y. Fu, V. Colonna, X. J. Mu, H. M. Kang *et al.,* 2013 Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science (New York, N.Y.) 342 (6154):1235587-1235587.

Kiiski, J. I., L. M. Pelttari, S. Khan, E. S. Freysteinsdottir, I. Reynisdottir *et al.,* 2014 Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. Proc Natl Acad Sci U S A 111 (42):15172-15177.

Kozarewa, I., Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman *et al.,* 2009 Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. Nat Methods 6 (4):291-295.

MacArthur, D. G., and C. Tyler-Smith, 2010 Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet 19 (R2):R125-130.

Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado, 2011 What can exome sequencing do for you? J Med Genet 48 (9):580-589.

Mardis, E. R., 2010 The $1,000 genome, the $100,000 analysis? Genome Med 2 (11):84-84.

Mascher, M., M. Jost, J.-E. Kuon, A. Himmelbach, A. Aßfalg *et al.,* 2014 Mapping-by-sequencing accelerates forward genetics in barley. Genome Biol 15 (6):R78-R78.

Mascher, M., T. A. Richmond, D. J. Gerhardt, A. Himmelbach, L. Clissold *et al.,* 2013 Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant J 76 (3):494-505.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.,* 2012 Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337 (6099):1190-1195.

McClure, M. C., D. Bickhart, D. Null, P. Vanraden, L. Xu *et al.,* 2014 Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. PLoS ONE 9 (3):e92769.

Meynert, A. M., M. Ansari, D. R. FitzPatrick, and M. S. Taylor, 2014 Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15 (1):247.

Mu, X. J., Z. J. Lu, Y. Kong, H. Y. K. Lam, and M. B. Gerstein, 2011 Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 39 (16):7058-7076.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham *et al.,* 2009 Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. Nature 461 (7261):272-276.

Pankin, A., C. Campoli, X. Dong, B. Kilian, R. Sharma *et al.,* 2014 Mapping-by-Sequencing Identifies HvPHYTOCHROME C as a Candidate Gene for the early maturity 5 Locus Modulating the Circadian Clock and Photoperiodic Flowering in Barley. Genetics 198 (1):383-396.

Rabbani, B., N. Mahdieh, K. Hosomichi, H. Nakaoka, and I. Inoue, 2012 Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 57 (10):621-632.

Robert, C., P. Fuentes-Utrilla, K. Troup, J. Loecherbach, F. Turner *et al.,* 2014 Design and development of exome capture sequencing for the domestic pig (Sus scrofa). BMC Genomics 15:9.

Roberts, R. J., M. O. Carneiro, and M. C. Schatz, 2013 The advantages of SMRT sequencing. Genome Biol 14 (7):405-405.

Sassi, C., R. Guerreiro, R. Gibbs, J. Ding, M. K. Lupton *et al.,* 2014 Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and p.S230R) in British early-onset Alzheimer's disease. Neurobiol Aging 35 (10):2422 e2413-2426.

Sboner, A., X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein, 2011 The real cost of sequencing: higher than you think! Genome Biol 12 (8):125-125.

Schick, U. M., P. L. Auer, J. C. Bis, H. Lin, P. Wei *et al.,* 2015 Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum Mol Genet 24 (2):559-571.

Shi, Y., Y. Li, D. Zhang, H. Zhang, Y. Li *et al.,* 2011 Exome Sequencing Identifies ZNF644 Mutations in High Myopia. PLoS Genet 7 (6):e1002084.

Shin, S. C., D. H. Ahn, S. J. Kim, H. Lee, T.-J. Oh *et al.,* 2013 Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS ONE 8 (7):e68824.

Smith, J. D., A. V. Hing, C. M. Clarke, N. M. Johnson, F. A. Perez *et al.,* 2014 Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontonasal dysostosis. Am J Hum Genet 95 (2):235-240.

Snape, K., E. Ruark, P. Tarpey, A. Renwick, C. Turnbull *et al.,* 2012 Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. Breast Cancer Res Treat 134 (1):429-433.

Stern, J., S. White, and K. Meurs, 2013 Extent of linkage disequilibrium in large-breed dogs: chromosomal and breed variation. Mammalian Genome 24 (9-10):409-415.

Sulonen, A. M., P. Ellonen, H. Almusa, M. Lepisto, S. Eldfors *et al.,* 2011 Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 12 (9):R94.
Taneri, B., E. Asilmaz, and T. Gaasterland, 2012 Biomedical impact of splicing mutations revealed through exome sequencing. Mol Med 18:314-319.
The 1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491 (7422):56-65.
The Deciphering Developmental Disorders, S., 2015 Large-scale discovery of novel genetic causes of developmental disorders. Nature 519 (7542):223-228.
The ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489 (7414):57-74.
The Fantom Consortium, The Riken Pmi, and Clst, 2014 A promoter-level mammalian expression atlas. Nature 507 (7493):462-470.
The International Barley Genome Sequencing Consortium, 2012 A physical, genetic and functional sequence assembly of the barley genome. Nature 491 (7426):711-716.
The International Wheat Genome Sequencing Consortium, 2014 A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345 (6194).
Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev *et al.,* 2006 The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science 313 (5793):1596-1604.
Vallender, E. J., 2011 Expanding whole exome resequencing into non-human primates. Genome Biol 12 (9):R87.
Veal, C., P. Freeman, K. Jacobs, O. Lancaster, S. Jamain *et al.,* 2012 A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics 13 (1):1-10.
Wang, H., X. Chen, L. Dudinsky, C. Patenia, Y. Chen *et al.,* 2011 Exome capture sequencing identifies a novel mutation in BBS4. Molecular Vision 17:3529-3540.
Ward, L. D., and M. Kellis, 2012 Interpreting non-coding variation in complex disease genetics. Nature biotechnology 30 (11):1095-1106.
Watson, M., 2014 Illuminating the future of DNA sequencing. Genome Biol 15 (2):108-108.
Wendler, N., M. Mascher, C. Noh, A. Himmelbach, U. Scholz *et al.,* 2014 Unlocking the secondary gene-pool of barley with next-generation sequencing. Plant Biotechnol J 12 (8):1122-1131.
Winfield, M. O., P. A. Wilkinson, A. M. Allen, G. L. Barker, J. A. Coghill *et al.,* 2012 Targeted re-sequencing of the allohexaploid wheat exome. Plant Biotechnol J 10 (6):733-742.
Worthey, E. A., A. N. Mayer, G. D. Syverson, D. Helbling, B. B. Bonacci *et al.,* 2011 Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13 (3):255-262.
Wright, C. F., T. W. Fitzgerald, W. D. Jones, S. Clayton, J. F. McRae *et al.,* 2015 Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet 385 (9975):1305-1314.
Xu, Y., B. Xiao, W. T. Jiang, L. Wang, H. Q. Gen *et al.,* 2014 A novel mutation identified in PKHD1 by targeted exome sequencing: guiding prenatal diagnosis for an ARPKD family. Gene 551 (1):33-38.
Yan, X. J., J. Xu, Z. H. Gu, C. M. Pan, G. Lu *et al.,* 2011 Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet 43 (4):309-315.
Zhou, L., and J. A. Holliday, 2012 Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics 13:703-703.
*Communicating editor: D. J. de Koning*

# BIBLIOGRAPHY

---

*"A mind needs books as a sword needs a whetstone, if it is to keep its edge. That's why I read so much, Jon Snow."*
-George R.R. Martin, A Game of Thrones

AARTSMA-RUS, A., GINJAAR, I. B. & BUSHBY, K. 2016. The importance of genetic diagnosis for Duchenne muscular dystrophy. *Journal of Medical Genetics*.

ABYZOV, A., URBAN, A. E., SNYDER, M. & GERSTEIN, M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research,* 21**,** 974-984.

ACS, P., BAUER, P. O., MAYER, B., BERA, T., MACALLISTER, R., MEZEY, E. & PASTAN, I. 2015. A novel form of ciliopathy underlies hyperphagia and obesity in Ankrd26 knockout mice. *Brain Struct Funct,* 220**,** 1511-28.

ADISSU, H. A., ESTABEL, J., SUNTER, D., TUCK, E., HOOKS, Y., CARRAGHER, D. M., CLARKE, K., KARP, N. A., PROJECT, S. M. G., NEWBIGGING, S., JONES, N., MORIKAWA, L., WHITE, J. K. & MCKERLIE, C. 2014. Histopathology reveals correlative and unique phenotypes in a high-throughput mouse phenotyping screen. *Disease Models & Mechanisms,* 7**,** 515-524.

ADZHUBEI, I., JORDAN, D. M. & SUNYAEV, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet,* Chapter 7**,** Unit7 20.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AGASHE, D., SANE, M., PHALNIKAR, K., DIWAN, G. D., HABIBULLAH, A., MARTINEZ-GOMEZ, N. C., SAHASRABUDDHE, V., POLACHEK, W., WANG, J., CHUBIZ, L. M. & MARX, C. J. 2016. Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. *Molecular Biology and Evolution,* 33**,** 1542-1553.

ÅGERSTAM, H., KARLSSON, C., HANSEN, N., SANDÉN, C., ASKMYR, M., VON PALFFY, S., HÖGBERG, C., RISSLER, M., WUNDERLICH, M., JULIUSSON, G., RICHTER, J., SJÖSTRÖM, K., BHATIA, R., MULLOY, J. C., JÄRÅS, M. & FIORETOS, T. 2015. Antibodies targeting human IL1RAP (IL1R3) show therapeutic effects in xenograft models of acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** 10786-10791.

AHONEN, S. J., ARUMILLI, M. & LOHI, H. 2013. A CNGB1 frameshift mutation in Papillon and Phalene dogs with progressive retinal atrophy. *PLoS One,* 8**,** e72122.

AI, H., HUANG, L. & REN, J. 2013. Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. *PLOS ONE,* 8**,** e56001.

AIRD, D., ROSS, M. G., CHEN, W.-S., DANIELSSON, M., FENNELL, T., RUSS, C., JAFFE, D. B., NUSBAUM, C. & GNIRKE, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology,* 12**,** R18-R18.

AKANNO, E. C., SCHENKEL, F. S., SARGOLZAEI, M., FRIENDSHIP, R. M. & ROBINSON, J. A. 2014. Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. *J Anim Breed Genet,* 131**,** 367-78.

ALFONSO, L. 2005. Use of meta-analysis to combine candidate gene association studies: application to study the relationship between the ESR PvuII polymorphism and sow litter size. *Genetics Selection Evolution,* 37**,** 417.

ALLEN, A. M., BARKER, G. L., WILKINSON, P., BURRIDGE, A., WINFIELD, M., COGHILL, J., UAUY, C., GRIFFITHS, S., JACK, P., BERRY, S., WERNER, P., MELICHAR, J. P., MCDOUGALL, J., GWILLIAM, R., ROBINSON, P. & EDWARDS, K. J. 2013. Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (Triticum aestivum L.). *Plant Biotechnol J,* 11**,** 279-95.

ALMAWI, W. Y., GUARINO, B. D., AL-SULAITI, M. A., AL-BUSAIDI, A. S., RACOUBIAN, E. & FINAN, R. R. 2013. Endothelial nitric oxide synthase gene variants and haplotypes associated with an increased risk of idiopathic recurrent miscarriage. *Hum Fertil (Camb),* 16**,** 200-6.

ALONSO-SPILSBURY, M. A., MOTA-ROJAS, D., MARTíNEZ-BURNES, J., ARCH, E., LÓPEZ MAYAGOITIA, A., RAMíREZ-NECOECHEA, R., OLMOS, A. & TRUJILLO, M. A. E. 2004. Use of oxytocin in penned sows and its effect on fetal intra-partum asphyxia. *Animal Reproduction Science,* 84**,** 157-167.

ALTMÄE, S., REIMAND, J., HOVATTA, O., ZHANG, P., KERE, J., LAISK, T., SAARE, M., PETERS, M., VILO, J., STAVREUS-EVERS, A. & SALUMETS, A. 2012. Research Resource: Interactome of Human Embryo Implantation: Identification of Gene Expression Pathways, Regulation, and Integrated Regulatory Networks. *Molecular Endocrinology,* 26**,** 203-217.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

ALVAREZ, G. 2008. Deviations from Hardy-Weinberg proportions for multiple alleles under viability selection. *Genet Res (Camb),* 90**,** 209-16.

AMARAL, A. J., MEGENS, H. J., CROOIJMANS, R. P., HEUVEN, H. C. & GROENEN, M. A. 2008. Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics,* 179**,** 569-79.

ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JORGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL,

C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MULLER, F., CONSORTIUM, T. F., FORREST, A. R. R., CARNINCI, P., REHLI, M. & SANDELIN, A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-461.

ANDREWS, S. 2010. FastQC: a quality control tool for high throughput sequence data.: Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

ANNUNEN, S., KÖRKKÖ, J., CZARNY, M., WARMAN, M. L., BRUNNER, H. G., KÄÄRIÄINEN, H., MULLIKEN, J. B., TRANEBJÆRG, L., BROOKS, D. G., COX, G. F., CRUYSBERG, J. R., CURTIS, M. A., DAVENPORT, S. L. H., FRIEDRICH, C. A., KAITILA, I., KRAWCZYNSKI, M. R., LATOS-BIELENSKA, A., MUKAI, S., OLSEN, B. R., SHINNO, N., SOMER, M., VIKKULA, M., ZLOTOGORA, J., PROCKOP, D. J. & ALA-KOKKO, L. 1999. Splicing Mutations of 54-bp Exons in the COL11A1 Gene Cause Marshall Syndrome, but Other Mutations Cause Overlapping Marshall/Stickler Phenotypes. *The American Journal of Human Genetics,* 65**,** 974-983.

ARCHIBALD, A. L., HALEY, C. S., BROWN, J. F., COUPERWHITE, S., MCQUEEN, H. A., NICHOLSON, D., COPPIETERS, W., VAN DE WEGHE, A., STRATIL, A., WINTERO, A. K. & ET AL. 1995. The PiGMaP consortium linkage map of the pig (Sus scrofa). *Mamm Genome,* 6**,** 157-75.

ARLOTTA, P., MOLYNEAUX, B. J., CHEN, J., INOUE, J., KOMINAMI, R. & MACKLIS, J. D. Neuronal Subtype-Specific Genes that Control Corticospinal Motor Neuron Development In Vivo. *Neuron,* 45**,** 207-221.

ARTERO-CASTRO, A., KONDOH, H., FERNÁNDEZ-MARCOS, P. J., SERRANO, M., Y CAJAL, S. R. & LLEONART, M. E. 2009. Rplp1 bypasses replicative senescence and contributes to transformation. *Experimental Cell Research,* 315**,** 1372-1383.

BADKE, Y. M., BATES, R. O., ERNST, C. W., SCHWAB, C. & STEIBEL, J. P. 2012. Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics,* 13**,** 24.

BAINBRIDGE, M. N., WANG, M., WU, Y., NEWSHAM, I., MUZNY, D. M., JEFFERIES, J. L., ALBERT, T. J., BURGESS, D. L. & GIBBS, R. A. 2011. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biology,* 12**,** R68-R68.

BAKER, M. 2012. De novo genome assembly: what every biologist should know. *Nat Meth,* 9**,** 333-337.

BAO, S., JIANG, R., KWAN, W., WANG, B., MA, X. & SONG, Y. Q. 2011. Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet,* 56**,** 406-14.

BARAS, A. 2018. *Large-Scale Sequencing in the UK Biobank to Facilitate Gene Discovery, Genome Sciences, and Precision Medicine | UK*

*Biobank* [Online]. Available: http://www.ukbiobank.ac.uk/2017/05/dr-aris-baras-large-scale-sequencing-in-the-uk-biobank-to-facilitate-gene-discovery-genome-sciences-and-precision-medicine [Accessed 14 May 2018].

BASHIR, A., KLAMMER, A. A., ROBINS, W. P., CHIN, C. S., WEBSTER, D., PAXINOS, E., HSU, D., ASHBY, M., WANG, S., PELUSO, P., SEBRA, R., SORENSON, J., BULLARD, J., YEN, J., VALDOVINO, M., MOLLOVA, E., LUONG, K., LIN, S., LAMAY, B., JOSHI, A., ROWE, L., FRACE, M., TARR, C. L., TURNSEK, M., DAVIS, B. M., KASARSKIS, A., MEKALANOS, J. J., WALDOR, M. K. & SCHADT, E. E. 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol,* 30.

BASSOLS, A., COSTA, C., ECKERSALL, P. D., OSADA, J., SABRIÀ, J. & TIBAU, J. 2014. The pig as an animal model for human pathologies: A proteomics perspective. *PROTEOMICS – Clinical Applications,* 8**,** 715-731.

BEAUDIN, A. E., PERRY, C. A., STABLER, S. P., ALLEN, R. H. & STOVER, P. J. 2012. Maternal Mthfd1 disruption impairs fetal growth but does not cause neural tube defects in mice. *The American Journal of Clinical Nutrition,* 95**,** 882-891.

BEAULIEU, A. D., AALHUS, J. L., WILLIAMS, N. H. & PATIENCE, J. F. 2010. Impact of piglet birth weight, birth order, and litter size on subsequent growth performance, carcass quality, muscle composition, and eating quality of pork. *J Anim Sci,* 88**,** 2767-78.

BECKMANN, J. S. 2015. Can We Afford to Sequence Every Newborn Baby's Genome? *Human Mutation,* 36**,** 283-286.

BELKADI, A., BOLZE, A., ITAN, Y., COBAT, A., VINCENT, Q. B., ANTIPENKO, A., SHANG, L., BOISSON, B., CASANOVA, J.-L. & ABEL, L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*.

BELTON, J.-M., MCCORD, R. P., GIBCUS, J., NAUMOVA, N., ZHAN, Y. & DEKKER, J. 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods (San Diego, Calif.),* 58**,** 10.1016/j.ymeth.2012.05.001.

BERLIN, K., KOREN, S., CHIN, C.-S., DRAKE, J. P., LANDOLIN, J. M. & PHILLIPPY, A. M. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotech,* 33**,** 623-630.

BIANCO, E., NEVADO, B., RAMOS-ONSINS, S. E. & PÉREZ-ENCISO, M. 2015. A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. *PLoS ONE,* 10**,** e0118867.

BICKHART, D. M., ROSEN, B. D., KOREN, S., SAYRE, B. L., HASTIE, A. R., CHAN, S., LEE, J., LAM, E. T., LIACHKO, I., SULLIVAN, S. T., BURTON, J. N., HUSON, H. J., NYSTROM, J. C., KELLEY, C. M., HUTCHISON, J. L., ZHOU, Y., SUN, J., CRISÀ, A., PONCE DE LEÓN, F. A., SCHWARTZ, J. C., HAMMOND, J. A., WALDBIESER, G. C., SCHROEDER, S. G., LIU, G. E., DUNHAM, M. J., SHENDURE,

J., SONSTEGARD, T. S., PHILLIPPY, A. M., VAN TASSELL, C. P. & SMITH, T. P. L. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics,* 49**,** 643.

BLANCKE, S., VAN BREUSEGEM, F., DE JAEGER, G., BRAECKMAN, J. & VAN MONTAGU, M. 2015. Fatal attraction: the intuitive appeal of GMO opposition. *Trends in Plant Science,* 20**,** 414-418.

BOBADILLA JOSEPH, L., MACEK, M., FINE JASON, P. & FARRELL PHILIP, M. 2002. Cystic fibrosis: A worldwide analysis of CFTR mutations—correlation with incidence data and application to screening. *Human Mutation,* 19**,** 575-606.

BOETZER, M., HENKEL, C. V., JANSEN, H. J., BUTLER, D. & PIROVANO, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* 27**,** 578-579.

BOLON, Y. T., HAUN, W. J., XU, W. W., GRANT, D., STACEY, M. G., NELSON, R. T., GERHARDT, D. J., JEDDELOH, J. A., STACEY, G., MUEHLBAUER, G. J., ORF, J. H., NAEVE, S. L., STUPAR, R. M. & VANCE, C. P. 2011. Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiol,* 156**,** 240-53.

BOLORMAA, S., GORE, K., VAN DER WERF, J. H., HAYES, B. J. & DAETWYLER, H. D. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Anim Genet,* 46**,** 544-56.

BOSI, E., DONATI, B., GALARDINI, M., BRUNETTI, S., SAGOT, M.-F., LIÓ, P., CRESCENZI, P., FANI, R. & FONDI, M. 2015. MeDuSa: a multi-draft based scaffolder. *Bioinformatics,* 31**,** 2443-2451.

BOSSE, M., MEGENS, H. J., MADSEN, O., PAUDEL, Y., FRANTZ, L. A. & SCHOOK, L. B. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet,* 8.

BOTSTEIN, D. & RISCH, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics,* 33**,** 228-237.

BOYLE, M. P. & DE BOECK, K. 2013. A new era in the treatment of cystic fibrosis: correction of the underlying CFTR defect. *The Lancet Respiratory Medicine,* 1**,** 158-163.

BRAS, J. M. & SINGLETON, A. B. 2011. Exome sequencing in Parkinson's disease. *Clin Genet,* 80**,** 104-9.

BROWN, C. G. Oxford Nanopore Technical Update. London Calling, 2018 London, UK.

BROWNING, B. L. & BROWNING, S. R. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics,* 84**,** 210-223.

BURBANO, H. A., HODGES, E., GREEN, R. E., BRIGGS, A. W., KRAUSE, J., MEYER, M., GOOD, J. M., MARICIC, T., JOHNSON, P. L., XUAN, Z., ROOKS, M., BHATTACHARJEE, A., BRIZUELA, L., ALBERT, F.

W., DE LA RASILLA, M., FORTEA, J., ROSAS, A., LACHMANN, M., HANNON, G. J. & PAABO, S. 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science,* 328**,** 723-5.

BURKARD, C., LILLICO, S. G., REID, E., JACKSON, B., MILEHAM, A. J., AIT-ALI, T., WHITELAW, C. B. A. & ARCHIBALD, A. L. 2017. Precision engineering for PRRSV resistance in pigs: Macrophages from genome edited pigs lacking CD163 SRCR5 domain are fully resistant to both PRRSV genotypes while maintaining biological function. *PLOS Pathogens,* 13**,** e1006206.

BURKARD, C., OPRIESSNIG, T., MILEHAM, A. J., STADEJEK, T., AIT-ALI, T., LILLICO, S. G., WHITELAW, C. B. A. & ARCHIBALD, A. L. 2018. Pigs lacking the scavenger receptor cysteine-rich domain 5 of CD163 are resistant to PRRSV-1 infection. *Journal of Virology*.

BURTON, G. J., YUNG, H. W., CINDROVA-DAVIES, T. & CHARNOCK-JONES, D. S. 2009. Placental Endoplasmic Reticulum Stress and Oxidative Stress in the Pathophysiology of Unexplained Intrauterine Growth Restriction and Early Onset Preeclampsia. *Placenta,* 30**,** 43-48.

BYUN, M., ABHYANKAR, A., LELARGE, V., PLANCOULAINE, S., PALANDUZ, A., TELHAN, L., BOISSON, B., PICARD, C., DEWELL, S., ZHAO, C., JOUANGUY, E., FESKE, S., ABEL, L. & CASANOVA, J. L. 2010. Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *J Exp Med,* 207**,** 2307-12.

CAI, J., LI, L., YE, L., JIANG, X., SHEN, L., GAO, Z., FANG, W., HUANG, F., SU, T., ZHOU, Y., WANG, W. & NING, G. 2015. Exome sequencing reveals mutant genes with low penetrance involved in MEN2A-associated tumorigenesis. *Endocr Relat Cancer,* 22**,** 23-33.

CALVERT, J. G., SLADE, D. E., SHIELDS, S. L., JOLIE, R., MANNAN, R. M., ANKENBAUER, R. G. & WELCH, S.-K. W. 2007. CD163 Expression Confers Susceptibility to Porcine Reproductive and Respiratory Syndrome Viruses. *Journal of Virology,* 81**,** 7371-7379.

CARLSON, D. F., LANCTO, C. A., ZANG, B., KIM, E.-S., WALTON, M., OLDESCHULTE, D., SEABURY, C., SONSTEGARD, T. S. & FAHRENKRUG, S. C. 2016. Production of hornless dairy cattle from genome-edited cell lines. *Nature Biotechnology,* 34**,** 479.

CARTEGNI, L., HASTINGS, M. L., CALARCO, J. A., STANCHINA, E. D. & KRAINER, A. R. 2006. Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2. *American Journal of Human Genetics,* 78**,** 63-77.

CASTRO-WALLACE, S. L., CHIU, C. Y., JOHN, K. K., STAHL, S. E., RUBINS, K. H., MCINTYRE, A. B. R., DWORKIN, J. P., LUPISELLA, M. L., SMITH, D. J., BOTKIN, D. J., STEPHENSON, T. A., JUUL, S., TURNER, D. J., IZQUIERDO, F., FEDERMAN, S., STRYKE, D., SOMASEKAR, S., ALEXANDER, N., YU, G., MASON, C. & BURTON, A. S. 2016. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *bioRxiv*.

CATENACCI, D. V. T., CERVANTES, G., YALA, S., NELSON, E. A., EL-HASHANI, E., KANTETI, R., EL DINALI, M., HASINA, R., BRÄGELMANN, J., SEIWERT, T., SANICOLA, M., HENDERSON, L., GRUSHKO, T. A., OLOPADE, O., KARRISON, T., BANG, Y.-J., KIM, W. H., TRETIAKOVA, M., VOKES, E., FRANK, D. A., KINDLER, H. L., HUET, H. & SALGIA, R. 2011. RON (MST1R) is a novel prognostic marker and therapeutic target for gastroesophageal adenocarcinoma. *Cancer Biology & Therapy,* 12**,** 9-46.

CETINKAYA, A., XIONG, JINGWEI R., VARGEL, İ., KÖSEMEHMETOĞLU, K., CANTER, HALIL İ., GERDAN, ÖMER F., LONGO, N., ALZAHRANI, A., CAMPS, MIREIA P., TASKIRAN, EKIM Z., LAUPHEIMER, S., BOTTO, LORENZO D., PARAMALINGAM, E., GORMEZ, Z., UZ, E., YUKSEL, B., RUACAN, Ş., SAĞıROĞLU, MAHMUT Ş., TAKAHASHI, T., REVERSADE, B. & AKARSU, NURTEN A. 2016. Loss-of-Function Mutations in ELMO2 Cause Intraosseous Vascular Malformation by Impeding RAC1 Signaling. *American Journal of Human Genetics,* 99**,** 299-317.

CHAISSON, M. J. & TESLER, G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics,* 13**,** 238.

CHAN, K. T., CREED, S. J. & BEAR, J. E. 2011. Unraveling the enigma: Progress towards understanding the Coronin family of actin regulators. *Trends in cell biology,* 21**,** 481-488.

CHANDRA, D., KORPI, E. R., MIRALLES, C. P., DE BLAS, A. L. & HOMANICS, G. E. 2005. GABAA receptor gamma 2 subunit knockdown mice have enhanced anxiety-like behavior but unaltered hypnotic response to benzodiazepines. *BMC Neurosci,* 6**,** 30.

CHELBI, S. T., MONDON, F., JAMMES, H., BUFFAT, C., MIGNOT, T.-M., TOST, J., BUSATO, F., GUT, I., REBOURCET, R., LAISSUE, P., TSATSARIS, V., GOFFINET, F., RIGOURD, V., CARBONNE, B., FERRÉ, F. & VAIMAN, D. 2007. Expressional and Epigenetic Alterations of Placental Serine Protease Inhibitors. *Hypertension,* 49**,** 76.

CHELBI, S. T., WILSON, M. L., VEILLARD, A.-C., INGLES, S. A., ZHANG, J., MONDON, F., GASCOIN-LACHAMBRE, G., DORIDOT, L., MIGNOT, T.-M., REBOURCET, R., CARBONNE, B., CONCORDET, J.-P., BARBAUX, S. & VAIMAN, D. 2012. Genetic and epigenetic mechanisms collaborate to control SERPINA3 expression and its association with placental diseases. *Human Molecular Genetics,* 21**,** 1968-1978.

CHEN, K., CHEN, L., FAN, X., WALLIS, J., DING, L. & WEINSTOCK, G. 2014. TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Research,* 24**,** 310-317.

CHEN, L., LIU, P., EVANS, T. C. & ETTWILLER, L. M. 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science,* 355**,** 752-756.

CHEN, M. C. 1989. Mapping Our Genes. Genome Projects: How Big, How Fast? Congress of the United States, Office of Technology Assessment. *The Yale Journal of Biology and Medicine,* 62**,** 235-236.
CHEN, M. H., CHEN, H., ZHOU, Z., RUAN, Y. C., WONG, H. Y., LU, Y. C., GUO, J. H., CHUNG, Y. W., HUANG, P. B., HUANG, H. F., ZHOU, W. L. & CHAN, H. C. 2010. Involvement of CFTR in oviductal HCO3− secretion and its effect on soluble adenylate cyclase-dependent early embryo development. *Human Reproduction,* 25**,** 1744-1754.
CHEN, S. H., WU, P.-S., CHOU, C.-H., YAN, Y.-T., LIU, H., WENG, S.-Y. & YANG-YEN, H.-F. 2007. A Knockout Mouse Approach Reveals that TCTP Functions as an Essential Factor for Cell Proliferation and Survival in a Tissue- or Cell Type–specific Manner. *Molecular Biology of the Cell,* 18**,** 2525-2532.
CHEN, Y.-C., LIU, T., YU, C.-H., CHIANG, T.-Y. & HWANG, C.-C. 2013. Effects of GC Bias in Next-Generation-Sequencing Data on <italic>De Novo</italic> Genome Assembly. *PLoS ONE,* 8**,** e62856.
CHIN, C.-S., PELUSO, P., SEDLAZECK, F. J., NATTESTAD, M., CONCEPCION, G. T., CLUM, A., DUNN, C., O'MALLEY, R., FIGUEROA-BALDERAS, R., MORALES-CRUZ, A., CRAMER, G. R., DELLEDONNE, M., LUO, C., ECKER, J. R., CANTU, D., RANK, D. R. & SCHATZ, M. C. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Meth,* 13**,** 1050-1054.
CHIN, C. S., ALEXANDER, D. H., MARKS, P., KLAMMER, A. A., DRAKE, J. & HEINER, C. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods,* 10.
CHOW, W., BRUGGER, K., CACCAMO, M., SEALY, I., TORRANCE, J. & HOWE, K. 2016. gEVAL — a web-based browser for evaluating genome assemblies. *Bioinformatics,* 32**,** 2508-2510.
CHRISTIANSEN, O. B. 1996. A fresh look at the causes and treatments of recurrent miscarriage, especially its immunological aspects. *Hum Reprod Update,* 2**,** 271-93.
CHRISTIANSON, W. T. 1992. Stillbirths, Mummies, Abortions, and Early Embryonic Death. *Veterinary Clinics of North America: Food Animal Practice,* 8**,** 623-639.
CHURCH, D. M., SCHNEIDER, V. A., GRAVES, T., AUGER, K., CUNNINGHAM, F. & BOUK, N. 2011. Modernizing reference genome assemblies. *PLoS Biol,* 9.
CHURCH, D. M., SCHNEIDER, V. A., STEINBERG, K. M., SCHATZ, M. C., QUINLAN, A. R., CHIN, C.-S., KITTS, P. A., AKEN, B., MARTH, G. T., HOFFMAN, M. M., HERRERO, J., MENDOZA, M. L. Z., DURBIN, R. & FLICEK, P. 2015. Extending reference assembly models. *Genome Biology,* 16**,** 13.
CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

CLAVERIE-MARTIN, F., GONZALEZ-PAREDES, F. J. & RAMOS-TRUJILLO, E. 2015. Splicing defects caused by exonic mutations in PKD1 as a new mechanism of pathogenesis in autosomal dominant polycystic kidney disease. *RNA Biology,* 12**,** 369-374.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., ACCINELLI, G. G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D. M., NAMAATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., PALMA, F. D., PHILLIPS, A., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*.

COCKETT, N. E., SMIT, M. A., BIDWELL, C. A., SEGERS, K., HADFIELD, T. L., SNOWDER, G. D., GEORGES, M. & CHARLIER, C. 2005. The callipyge mutation and other genes that affect muscle hypertrophy in sheep. *Genet Sel Evol,* 37 Suppl 1**,** S65-81.

COMPEAU, P. E. C., PEVZNER, P. A. & TESLER, G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotech,* 29**,** 987-991.

CORBIN, L. J., KRANIS, A., BLOTT, S. C., SWINBURNE, J. E., VAUDIN, M., BISHOP, S. C. & WOOLLIAMS, J. A. 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. *Genet Sel Evol,* 46**,** 9.

CORNISH, A. & GUDA, C. 2015. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International,* 2015**,** 11.

COSART, T., BEJA-PEREIRA, A., CHEN, S., NG, S. B., SHENDURE, J. & LUIKART, G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics,* 12**,** 347.

COZLER, Y. L., GUYOMARC'H, C., PICHODO, X., QUINIO, P.-Y. & PELLOIS, H. 2002. Factors associated with stillborn and mummified piglets in high-prolific sows. *Anim. Res.,* 51**,** 261-268.

CRAIG, T. J., SHIMOMURA, K., HOLL, R. W., FLANAGAN, S. E., ELLARD, S. & ASHCROFT, F. M. 2009. An In-Frame Deletion in Kir6.2 (KCNJ11) Causing Neonatal Diabetes Reveals a Site of Interaction between Kir6.2 and SUR1. *The Journal of Clinical Endocrinology & Metabolism,* 94**,** 2551-2557.

CRETU STANCU, M., VAN ROOSMALEN, M. J., RENKENS, I., NIEBOER, M. M., MIDDELKAMP, S., DE LIGT, J., PREGNO, G., GIACHINO, D., MANDRILE, G., ESPEJO VALLE-INCLAN, J., KORZELIUS, J., DE BRUIJN, E., CUPPEN, E., TALKOWSKI, M. E., MARSCHALL, T., DE RIDDER, J. & KLOOSTERMAN, W. P. 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications,* 8**,** 1326.

CUTTING, G. R. 2014. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics,* 16**,** 45.
DAETWYLER, H. D., CAPITAN, A., PAUSCH, H., STOTHARD, P., VAN BINSBERGEN, R., BRØNDUM, R. F., LIAO, X., DJARI, A., RODRIGUEZ, S. C., GROHS, C., ESQUERRÉ, D., BOUCHEZ, O., ROSSIGNOL, M.-N., KLOPP, C., ROCHA, D., FRITZ, S., EGGEN, A., BOWMAN, P. J., COOTE, D., CHAMBERLAIN, A. J., ANDERSON, C., VANTASSELL, C. P., HULSEGGE, I., GODDARD, M. E., GULDBRANDTSEN, B., LUND, M. S., VEERKAMP, R. F., BOICHARD, D. A., FRIES, R. & HAYES, B. J. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics,* 46**,** 858.
DALL'OLIO, S., FONTANESI, L., TOGNAZZI, L., BUTTAZZONI, L., GALLO, M. & RUSSO, V. 2011. ESR1 and ESR2 gene markers are not associated with number of piglets born alive in Italian Large White sows. *Italian Journal of Animal Science,* 10**,** e35.
DAMAS, J., O'CONNOR, R., FARRÉ, M., LENIS, V. P. E., MARTELL, H. J., MANDAWALA, A., FOWLER, K., JOSEPH, S., SWAIN, M. T., GRIFFIN, D. K. & LARKIN, D. M. 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research,* 27**,** 875-884.
DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E. & DEPRISTO, M. A. 2011. The variant call format and VCFtools. *Bioinformatics,* 27.
DAS, S. K., AUSTIN, M. D., AKANA, M. C., DESHPANDE, P., CAO, H. & XIAO, M. 2010. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research,* 38**,** e177-e177.
DAVEY, J. W., CHOUTEAU, M., BARKER, S. L., MAROJA, L., BAXTER, S. W., SIMPSON, F., MERRILL, R. M., JORON, M., MALLET, J., DASMAHAPATRA, K. K. & JIGGINS, C. D. 2016. Major Improvements to the <em>Heliconius melpomene</em> Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. *G3: Genes|Genomes|Genetics,* 6**,** 695.
DAYARIAN, A., MICHAEL, T. P. & SENGUPTA, A. M. 2010. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics,* 11**,** 345-345.
DE BOECK, K., ZOLIN, A., CUPPENS, H., OLESEN, H. V. & VIVIANI, L. 2014. The relative frequency of CFTR mutation classes in European patients with cystic fibrosis. *Journal of Cystic Fibrosis,* 13**,** 403-409.
DE DONATO, M., PETERS, S. O., MITCHELL, S. E., HUSSAIN, T. & IMUMORIN, I. G. 2013. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One,* 8**,** e62137.
DE KONING, A. P. J., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics,* 7**,** e1002384.

DERKS, M. F. L., LOPES, M. S., BOSSE, M., MADSEN, O., DIBBITS, B., HARLIZIUS, B., GROENEN, M. A. M. & MEGENS, H.-J. 2018. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *PLOS Genetics,* 14**,** e1007661.

DERKS, M. F. L., MEGENS, H.-J., BOSSE, M., LOPES, M. S., HARLIZIUS, B. & GROENEN, M. A. M. 2017. A systematic survey to identify lethal recessive variation in highly managed pig populations. *BMC Genomics,* 18**,** 858.

DINARELLO, C. A. 2000. Proinflammatory cytokines. *Chest,* 118**,** 503-8.

DIVINA, P., KVITKOVICOVA, A., BURATTI, E. & VORECHOVSKY, I. 2009. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *European Journal Of Human Genetics,* 17**,** 759.

DOLEZALOVA, D., HRUSKA-PLOCHAN, M., BJARKAM, C. R., SORENSEN, J. C., CUNNINGHAM, M., WEINGARTEN, D., CIACCI, J. D., JUHAS, S., JUHASOVA, J., MOTLIK, J., HEFFERAN, M. P., HAZEL, T., JOHE, K., CARROMEU, C., MUOTRI, A., BUI, J., STRNADEL, J. & MARSALA, M. 2014. Pig models of neurodegenerative disorders: Utilization in cell replacement-based preclinical safety and efficacy studies. *J Comp Neurol,* 522**,** 2784-801.

DONMEZ, N. & BRUDNO, M. 2013. SCARPA: scaffolding reads with practical algorithms. *Bioinformatics,* 29**,** 428-434.

DRAKE, J. SMRT Link and Analysis Tools for PacBio data. SMRTLeiden, 2018 NL.

DRMANAC, R. 2011. The advent of personal genome sequencing. *Genet Med,* 13**,** 188-190.

DUBEY, J. P. 1999. Recent advances in Neospora and neosporosis. *Veterinary Parasitology,* 84**,** 349-367.

DULBECCO, R. 1986. A turning point in cancer research: sequencing the human genome. *Science,* 231**,** 1055-6.

EDWARDS, A., SOARES, A., RASSNER, S., GREEN, P., FELIX, J. & MITCHELL, A. 2017. Deep Sequencing: Intra-Terrestrial Metagenomics Illustrates The Potential Of Off-Grid Nanopore DNA Sequencing. *bioRxiv*.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P., BETTMAN, B., BIBILLO, A., BJORNSON, K., CHAUDHURI, B., CHRISTIANS, F., CICERO, R., CLARK, S., DALAL, R., DEWINTER, A., DIXON, J., FOQUET, M., GAERTNER, A., HARDENBOL, P., HEINER, C., HESTER, K., HOLDEN, D., KEARNS, G., KONG, X., KUSE, R., LACROIX, Y., LIN, S., LUNDQUIST, P., MA, C., MARKS, P., MAXHAM, M., MURPHY, D., PARK, I., PHAM, T., PHILLIPS, M., ROY, J., SEBRA, R., SHEN, G., SORENSON, J., TOMANEY, A., TRAVERS, K., TRULSON, M., VIECELI, J., WEGENER, J., WU, D., YANG, A., ZACCARIN, D., ZHAO, P., ZHONG, F., KORLACH, J. & TURNER, S. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323**,** 133-138.

EKBLOM, R. & WOLF, J. B. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl,* 7**,** 1026-42.
ENGLISH, A. C., RICHARDS, S., HAN, Y., WANG, M., VEE, V., QU, J., QIN, X., MUZNY, D. M., REID, J. G., WORLEY, K. C. & GIBBS, R. A. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE,* 7**,** e47768.
ENGLISH, A. C., SALERNO, W. J. & REID, J. G. 2014. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics,* 15**,** 180.
ERIKSON, G. A., BODIAN, D. L., RUEDA, M., MOLPARIA, B., SCOTT, E. R., SCOTT-VAN ZEELAND, A. A., TOPOL, S. E., WINEINGER, N. E., NIEDERHUBER, J. E., TOPOL, E. J. & TORKAMANI, A. 2016. Whole Genome Sequencing of a Healthy Aging Cohort. *Cell,* 165**,** 1002-1011.
EVANI, U. S., CHALLIS, D., YU, J., JACKSON, A. R., PAITHANKAR, S., BAINBRIDGE, M. N., JAKKAMSETTI, A., PHAM, P., COARFA, C., MILOSAVLJEVIC, A. & YU, F. 2012. Atlas2 Cloud: a framework for personal genome analysis in the cloud. *BMC Genomics,* 13**,** S19.
EXOME AGGREGATION CONSORTIUM (EXAC). 2014. Cambridge, MA. Available: http://exac.broadinstitute.org [Accessed 04/2015].
FAJARDO, K. V. F., ADAMS, D., PROGRAM, N. C. S., MASON, C. E., SINCAN, M., TIFFT, C., TORO, C., BOERKOEL, C. F., GAHL, W. & MARKELLO, T. 2012. Detecting false positive signals in exome sequencing. *Human Mutation,* 33**,** 609-613.
FAN, N. & LAI, L. 2013. Genetically modified pig models for human diseases. *J Genet Genomics,* 40**,** 67-73.
FANG, X., MOU, Y., HUANG, Z., LI, Y., HAN, L., ZHANG, Y., FENG, Y., CHEN, Y., JIANG, X., ZHAO, W., SUN, X., XIONG, Z., YANG, L., LIU, H., FAN, D., MAO, L., REN, L., LIU, C., WANG, J., LI, K., WANG, G., YANG, S., LAI, L., ZHANG, G., LI, Y., WANG, J., BOLUND, L., YANG, H., WANG, J., FENG, S., LI, S. & DU, Y. 2012. The sequence and analysis of a Chinese pig genome. *GigaScience,* 1**,** 16-16.
FONTANESI, L., TAZZOLI, M., BERETTI, F. & RUSSO, V. 2006. Mutations in the melanocortin 1 receptor (MC1R) gene are associated with coat colours in the domestic rabbit (Oryctolagus cuniculus). *Animal Genetics,* 37**,** 489-493.
FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS 2015. FAO GEONETWORK. Rome, Italy: FAO.
FUKUDA, M. N., SUGIHARA, K. & NAKAYAMA, J. 2008. Trophinin: what embryo implantation teaches us about human cancer. *Cancer Biol Ther,* 7**,** 1165.
FULLER, C. W., MIDDENDORF, L. R., BENNER, S. A., CHURCH, G. M., HARRIS, T., HUANG, X., JOVANOVICH, S. B., NELSON, J. R., SCHLOSS, J. A., SCHWARTZ, D. C. & VEZENOV, D. V. 2009. The challenges of sequencing by synthesis. *Nat Biotech,* 27**,** 1013-1023.
GAILANI, M. R. & BALE, A. E. 1997. Developmental Genes and Cancer: Role of Patched in Basal Cell Carcinoma of the Skin. *JNCI: Journal of the National Cancer Institute,* 89**,** 1103-1109.

GALARDINI, M., BIONDI, E. G., BAZZICALUPO, M. & MENGONI, A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine,* 6**,** 11-11.
GALLEGO-BUSTOS, F., GOTEA, V., RAMOS-AMADOR, J. T., RODRÍGUEZ-PENA, R., GIL-HERRERA, J., SASTRE, A., DELMIRO, A., RAI, G., ELNITSKI, L., GONZÁLEZ-GRANADO, L. I. & ALLENDE, L. M. 2016. A Case of IL-7R Deficiency Caused by a Novel Synonymous Mutation and Implications for Mutation Screening in SCID Diagnosis. *Frontiers in Immunology,* 7.
GALLENBERGER, M., MEINEL, D. M., KROEBER, M., WEGNER, M., MILKEREIT, P., BÖSL, M. R. & TAMM, E. R. 2011. Lack of WDR36 leads to preimplantation embryonic lethality in mice and delays the formation of small subunit ribosomal RNA in human cells in vitro. *Human Molecular Genetics,* 20**,** 422-435.
GAO, S., SUNG, W.-K. & NAGARAJAN, N. 2011. Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences. *Journal of Computational Biology,* 18**,** 1681-1691.
GARBUS, I., ROMERO, J. R., VALARIK, M., VANŽUROVÁ, H., KARAFIÁTOVÁ, M., CÁCCAMO, M., DOLEŽEL, J., TRANQUILLI, G., HELGUERA, M. & ECHENIQUE, V. 2015. Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. *BMC Genomics,* 16**,** 375.
GARRISON, E. & MARTH, G. 2012. *Haplotype-based variant detection from short-read sequencing*.
GECZ, J. & CORBETT, M. 2015. Developmental disorders: deciphering exomes on a grand scale. *The Lancet,* 385**,** 1266-1267.
GHANEM, M. E., NAKAO, T. & NISHIBORI, M. 2006. Deficiency of uridine monophosphate synthase (DUMPS) and X-chromosome deletion in fetal mummification in cattle. *Animal Reproduction Science,* 91**,** 45-54.
GHANEM, M. E., NISHIBORI, M., NAKAO, T. & MORIYOSHI, M. 2005. DNA Extraction from Bovine Mummified Fetuses and Detection of Factor XI Gene Deficiency in the Mummies. *Journal of Reproduction and Development,* 51**,** 347-352.
GILLET-MARKOWSKA, A., RICHARD, H., FISCHER, G. & LAFONTAINE, I. 2015. Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics,* 31**,** 801-808.
GILLY, A., SUVEGES, D., KUCHENBAECKER, K., POLLARD, M. O., SOUTHAM, L., HATZIKOTOULAS, K., FARMAKI, A.-E., BJORNLAND, T., WAPLES, R., APPEL, E. V. R., CASALONE, E., MELLONI, G., KILIAN, B., RAYNER, N. W., NTALLA, I., KUNDU, K., WALTER, K., DANESH, J., BUTTERWORTH, A., BARROSO, I., TSAFANTAKIS, E., DEDOUSSIS, G., MOLTKE, I. & ZEGGINI, E. 2018. Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *bioRxiv*.
GODDARD, M. E., KEMPER, K. E., MACLEOD, I. M., CHAMBERLAIN, A. J. & HAYES, B. J. 2016. Genetics of complex traits: prediction of

phenotype, identification of causal polymorphisms and genetic architecture. *Proc Biol Sci,* 283.
GONEN, S., ROS-FREIXEDES, R., BATTAGIN, M., GORJANC, G. & HICKEY, J. M. 2017. A method for the allocation of sequencing resources in genotyped livestock populations. *Genetics, selection, evolution : GSE* [Online], 49. Available: https://doi.org/10.1186/s12711-017-0322-5 [Accessed 2017/05//].
GONG, L., WONG, C.-H., CHENG, W.-C., TJONG, H., MENGHI, F., NGAN, C. Y., LIU, E. T. & WEI, C.-L. 2018. Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nature Methods*.
GORJANC, G., CLEVELAND, M. A., HOUSTON, R. D. & HICKEY, J. M. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genetics, selection, evolution : GSE* [Online], 47. Available: http://europepmc.org/abstract/MED/25887531 [Accessed 2015/03//].
GREIF, P. A., DUFOUR, A., KONSTANDIN, N. P., KSIENZYK, B., ZELLMEIER, E., TIZAZU, B., STURM, J., BENTHAUS, T., HEROLD, T., YAGHMAIE, M., DORGE, P., HOPFNER, K. P., HAUSER, A., GRAF, A., KREBS, S., BLUM, H., KAKADIA, P. M., SCHNEIDER, S., HOSTER, E., SCHNEIDER, F., STANULLA, M., BRAESS, J., SAUERLAND, M. C., BERDEL, W. E., BUCHNER, T., WOERMANN, B. J., HIDDEMANN, W., SPIEKERMANN, K. & BOHLANDER, S. K. 2012. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood,* 120**,** 395-403.
GRIFFIN, J. 2013. Methods of Sperm DNA Extraction for Genetic and Epigenetic Studies. *In:* CARRELL, D. T. & ASTON, K. I. (eds.) *Spermatogenesis: Methods and Protocols.* Totowa, NJ: Humana Press.
GRIFFITH, A. J., SPRUNGER, L. K., SIRKO-OSADSA, D. A., TILLER, G. E., MEISLER, M. H. & WARMAN, M. L. 1998. Marshall Syndrome Associated with a Splicing Defect at the COL11A1 Locus. *The American Journal of Human Genetics,* 62**,** 816-823.
GRIFFITHS, A. J. F., WESSLER, S. R., CARROLL, S. B. & DOEBLEY, J. 2015. *Introduction to genetic analysis*, Eleventh edition. New York, NY : W.H. Freeman Company, 2015.
GRISENDI, S., BERNARDI, R., ROSSI, M., CHENG, K., KHANDKER, L., MANOVA, K. & PANDOLFI, P. P. 2005. Role of nucleophosmin in embryonic development and tumorigenesis. *Nature,* 437**,** 147.
GRITSENKO, A. A., NIJKAMP, J. F., REINDERS, M. J. & DE RIDDER, D. 2012. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics,* 28.
GROENEN, M. A., ARCHIBALD, A. L., UENISHI, H., TUGGLE, C. K., TAKEUCHI, Y., ROTHSCHILD, M. F., ROGEL-GAILLARD, C., PARK, C., MILAN, D., MEGENS, H. J., LI, S., LARKIN, D. M., KIM, H., FRANTZ, L. A., CACCAMO, M., AHN, H., AKEN, B. L., ANSELMO, A., ANTHON, C., AUVIL, L., BADAOUI, B., BEATTIE, C. W., BENDIXEN,

C., BERMAN, D., BLECHA, F., BLOMBERG, J., BOLUND, L., BOSSE, M., BOTTI, S., BUJIE, Z., BYSTROM, M., CAPITANU, B., CARVALHO-SILVA, D., CHARDON, P., CHEN, C., CHENG, R., CHOI, S. H., CHOW, W., CLARK, R. C., CLEE, C., CROOIJMANS, R. P., DAWSON, H. D., DEHAIS, P., DE SAPIO, F., DIBBITS, B., DROU, N., DU, Z. Q., EVERSOLE, K., FADISTA, J., FAIRLEY, S., FARAUT, T., FAULKNER, G. J., FOWLER, K. E., FREDHOLM, M., FRITZ, E., GILBERT, J. G., GIUFFRA, E., GORODKIN, J., GRIFFIN, D. K., HARROW, J. L., HAYWARD, A., HOWE, K., HU, Z. L., HUMPHRAY, S. J., HUNT, T., HORNSHOJ, H., JEON, J. T., JERN, P., JONES, M., JURKA, J., KANAMORI, H., KAPETANOVIC, R., KIM, J., KIM, J. H., KIM, K. W., KIM, T. H., LARSON, G., LEE, K., LEE, K. T., LEGGETT, R., LEWIN, H. A., LI, Y., LIU, W., LOVELAND, J. E., LU, Y., LUNNEY, J. K., MA, J., MADSEN, O., MANN, K., MATTHEWS, L., MCLAREN, S., MOROZUMI, T., MURTAUGH, M. P., NARAYAN, J., NGUYEN, D. T., NI, P., OH, S. J., ONTERU, S., PANITZ, F., PARK, E. W., et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* 491**,** 393-8.

GROENEN, M. A. M. 2016. A decade of pig genome sequencing: a window on pig domestication and evolution. *Genetics Selection Evolution,* 48**,** 23.

GROSSMANN, V., TIACCI, E., HOLMES, A. B., KOHLMANN, A., MARTELLI, M. P., KERN, W., SPANHOL-ROSSETO, A., KLEIN, H. U., DUGAS, M., SCHINDELA, S., TRIFONOV, V., SCHNITTGER, S., HAFERLACH, C., BASSAN, R., WELLS, V. A., SPINELLI, O., CHAN, J., ROSSI, R., BALDONI, S., DE CAROLIS, L., GOETZE, K., SERVE, H., PECENY, R., KREUZER, K. A., ORUZIO, D., SPECCHIA, G., DI RAIMONDO, F., FABBIANO, F., SBORGIA, M., LISO, A., FARINELLI, L., RAMBALDI, A., PASQUALUCCI, L., RABADAN, R., HAFERLACH, T. & FALINI, B. 2011. Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood,* 118**,** 6153-63.

GUAN, P. & SUNG, W.-K. 2016. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods,* 102**,** 36-49.

GUNDERSEN, S., KALAŠ, M., ABUL, O., FRIGESSI, A., HOVIG, E. & SANDVE, G. K. 2011. Identifying elemental genomic track types and representing them uniformly. *BMC Bioinformatics,* 12**,** 494.

HABIER, D., FERNANDO, R. L. & DEKKERS, J. C. M. 2009. Genomic Selection Using Low-Density Marker Panels. *Genetics,* 182**,** 343.

HAFEZ, A., SQUIRES, R., PEDRACINI, A., JOSHI, A., SEEGMILLER, E. R. & OXFORD, T. J. 2015. Col11a1 Regulates Bone Microarchitecture during Embryonic Development. *Journal of Developmental Biology,* 3.

HAMPTON, O. A., ENGLISH, A. C., WANG, M., SALERNO, W. J., LIU, Y., MUZNY, D. M., HAN, Y., WHEELER, D. A., WORLEY, K. C., LUPSKI, J. R. & GIBBS, R. A. 2017. SVachra: a tool to identify genomic structural variation in mate pair sequencing data containing inward and outward facing reads. *BMC Genomics,* 18**,** 691.

HAN, K., CHEN, H., GENNARINO, V. A., RICHMAN, R., LU, H.-C. & ZOGHBI, H. Y. 2015. Fragile X-like behaviors and abnormal cortical dendritic spines in Cytoplasmic FMR1-interacting protein 2-mutant mice. *Human Molecular Genetics,* 24**,** 1813-1823.
HASUWA, H., UEDA, J., IKAWA, M. & OKABE, M. 2013. MiR-200b and miR-429 Function in Mouse Ovulation and Are Essential for Female Fertility. *Science,* 341**,** 71.
HATANO, S., KIMATA, K., HIRAIWA, N., KUSAKABE, M., ISOGAI, Z., ADACHI, E., SHINOMURA, T. & WATANABE, H. 2012. Versican/PG-M is essential for ventricular septal formation subsequent to cardiac atrioventricular cushion development. *Glycobiology,* 22**,** 1268-1277.
HATEM, A., BOZDAĞ, D., TOLAND, A. E. & ÇATALYÜREK, Ü. V. 2013. Benchmarking short sequence mapping tools. *BMC Bioinformatics,* 14**,** 184-184.
HAUBOLD, B. & WIEHE, T. 2006. How repetitive are genomes? *BMC Bioinformatics,* 7**,** 541.
HAUN, W. J., HYTEN, D. L., XU, W. W., GERHARDT, D. J., ALBERT, T. J., RICHMOND, T., JEDDELOH, J. A., JIA, G., SPRINGER, N. M., VANCE, C. P. & STUPAR, R. M. 2011. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol,* 155**,** 645-55.
HAYES, M., PYON, Y. S. & LI, J. 2012. A Model-Based Clustering Method for Genomic Structural Variant Prediction and Genotyping Using Paired-End Sequencing Data. *PLOS ONE,* 7**,** e52881.
HEINRICH, V., KAMPHANS, T., STANGE, J., PARKHOMCHUK, D., HECHT, J., DICKHAUS, T., ROBINSON, P. N. & KRAWITZ, P. M. 2013. Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine,* 5**,** 69-69.
HEIT, C., JACKSON, B. C., MCANDREWS, M., WRIGHT, M. W., THOMPSON, D. C., SILVERMAN, G. A., NEBERT, D. W. & VASILIOU, V. 2013. Update of the human and mouse SERPIN gene superfamily. *Human Genomics,* 7**,** 22-22.
HENDERSON, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics***,** 423-447.
HENRY, I. M., NAGALAKSHMI, U., LIEBERMAN, M. C., NGO, K. J., KRASILEVA, K. V., VASQUEZ-GROSS, H., AKHUNOVA, A., AKHUNOV, E., DUBCOVSKY, J., TAI, T. H. & COMAI, L. 2014. Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. *Plant Cell,* 26**,** 1382-1397.
HERCEG, Z., HULLA, W., GELL, D., CUENIN, C., LLEONART, M., JACKSON, S. & WANG, Z.-Q. 2001. Disruption of Trrap causes early embryonic lethality and defects in cell cycle progression. *Nature Genetics,* 29**,** 206.
HERNANDEZ, D., FRANÇOIS, P., FARINELLI, L., ØSTERÅS, M. & SCHRENZEL, J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research,* 18**,** 802-809.

HERNANDEZ, S. C., FINLAYSON, H. A., ASHWORTH, C. J., HALEY, C. S. & ARCHIBALD, A. L. 2014. A genome-wide linkage analysis for reproductive traits in F(2) Large White × Meishan cross gilts. *Animal Genetics,* 45**,** 191-197.

HICKEY, J. M. 2013. Sequencing millions of animals for genomic selection 2.0. *Journal of Animal Breeding and Genetics,* 130**,** 331-332.

HILL, W. G. 2014. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics,* 196**,** 1-16.

HOENEN, T., GROSETH, A., ROSENKE, K., FISCHER, R. J., HOENEN, A., JUDSON, S. D., MARTELLARO, C., FALZARANO, D., MARZI, A., SQUIRES, R. B., WOLLENBERG, K. R., DE WIT, E., PRESCOTT, J., SAFRONETZ, D., VAN DOREMALEN, N., BUSHMAKER, T., FELDMANN, F., MCNALLY, K., BOLAY, F. K., FIELDS, B., SEALY, T., RAYFIELD, M., NICHOL, S. T., ZOON, K. C., MASSAQUOI, M., MUNSTER, V. J. & FELDMANN, H. 2016. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerging Infectious Diseases,* 22**,** 331-334.

HOLM, B., BAKKEN, M., VANGEN, O. & REKAYA, R. 2004. Genetic analysis of litter size, parturition length, and birth assistance requirements in primiparous sows using a joint linear-threshold animal model. *J Anim Sci,* 82**,** 2528-33.

HONG, E. P. & PARK, J. W. 2012. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics,* 10**,** 117-122.

HORNER, D. S., PAVESI, G., CASTRIGNANO, T., DE MEO, P. D., LIUNI, S., SAMMETH, M., PICARDI, E. & PESOLE, G. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform,* 11**,** 181-97.

HOWARD, D. M., PONG-WONG, R., KNAP, P. W. & WOOLLIAMS, J. A. 2017. Use of haplotypes to identify regions harbouring lethal recessive variants in pigs. *Genetics Selection Evolution,* 49**,** 57.

HRYHOROWICZ, M., ZEYLAND, J., SŁOMSKI, R. & LIPIŃSKI, D. 2017. Genetically Modified Pigs as Organ Donors for Xenotransplantation. *Molecular Biotechnology,* 59**,** 435-444.

HU, H., HUFF, C. D., MOORE, B., FLYGARE, S., REESE, M. G. & YANDELL, M. 2013. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol,* 37**,** 622-34.

HUMPHRAY, S., SCOTT, C., CLARK, R., MARRON, B., BENDER, C., CAMM, N., DAVIS, J., JENKS, A., NOON, A., PATEL, M., SEHRA, H., YANG, F., ROGATCHEVA, M., MILAN, D., CHARDON, P., ROHRER, G., NONNEMAN, D., DE JONG, P., MEYERS, S., ARCHIBALD, A., BEEVER, J., SCHOOK, L. & ROGERS, J. 2007. A high utility integrated map of the pig genome. *Genome Biol,* 8.

HUNT, M., NEWBOLD, C., BERRIMAN, M. & OTTO, T. 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome Biol,* 15.

HWANG, S., KIM, E., LEE, I. & MARCOTTE, E. M. 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. 5**,** 17875.
IGLESIAS, A., ANYANE-YEBOA, K., WYNN, J., WILSON, A., TRUITT CHO, M., GUZMAN, E., SISSON, R., EGAN, C. & CHUNG, W. K. 2014. The usefulness of whole-exome sequencing in routine clinical practice. *Genet Med,* 16**,** 922-31.
INTERNATIONAL HUMAN GENOME SEQUENCING, C. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860.
INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.
ISHII, T. & ARAKI, M. 2016. Consumer acceptance of food crops developed by genome editing. *Plant Cell Reports,* 35**,** 1507-1518.
JAIN, M., KOREN, S., MIGA, K. H., QUICK, J., RAND, A. C., SASANI, T. A., TYSON, J. R., BEGGS, A. D., DILTHEY, A. T., FIDDES, I. T., MALLA, S., MARRIOTT, H., NIETO, T., O'GRADY, J., OLSEN, H. E., PEDERSEN, B. S., RHIE, A., RICHARDSON, H., QUINLAN, A. R., SNUTCH, T. P., TEE, L., PATEN, B., PHILLIPPY, A. M., SIMPSON, J. T., LOMAN, N. J. & LOOSE, M. 2018a. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*.
JAIN, M., OLSEN, H. E., TURNER, D. J., STODDART, D., BULAZEL, K. V., PATEN, B., HAUSSLER, D., WILLARD, H. F., AKESON, M. & MIGA, K. H. 2018b. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*.
JANAWAY, R. C., WILSON, A. S., DÍAZ, G. C. & GUILLEN, S. 2009. Taphonomic Changes to the Buried Body in Arid Environments: An Experimental Case Study in Peru. *In:* RITZ, K., DAWSON, L. & MILLER, D. (eds.) *Criminal and Environmental Soil Forensics.* Dordrecht: Springer Netherlands.
JANSEN, H. J., LIEM, M., JONG-RAADSEN, S. A., DUFOUR, S., WELTZIEN, F.-A., SWINKELS, W., KOELEWIJN, A., PALSTRA, A. P., PELSTER, B., SPAINK, H. P., VAN DEN THILLART, G. E., DIRKS, R. P. & HENKEL, C. V. 2017. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *bioRxiv*.
JAROUDI, S. & SENGUPTA, S. 2007. DNA repair in mammalian embryos. *Mutation Research/Reviews in Mutation Research,* 635**,** 53-77.
JIAN, X., BOERWINKLE, E. & LIU, X. 2013. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genetics In Medicine,* 16**,** 497.
JIANG, Y., WANG, Y. & BRUDNO, M. 2012. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics,* 28**,** 2576-2583.
JIAO, W.-B. & SCHNEEBERGER, K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology,* 36**,** 64-70.

JIMENEZ, P. T., MAINIGI, M. A., WORD, R. A., KRAUS, W. L. & MENDELSON, C. R. 2016. miR-200 Regulates Endometrial Development During Early Pregnancy. *Molecular Endocrinology,* 30**,** 977-987.

JOHANSSON, H. & SIMONSSON, S. 2010. Core transcription factors, Oct4, Sox2 and Nanog, individually form complexes with nucleophosmin (Npm1) to control embryonic stem (ES) cell fate determination. *Aging (Albany NY),* 2**,** 815-822.

JOHANSSON, S., IRGENS, H., CHUDASAMA, K. K., MOLNES, J., AERTS, J., ROQUE, F. S., JONASSEN, I., LEVY, S., LIMA, K., KNAPPSKOG, P. M., BELL, G. I., MOLVEN, A. & NJOLSTAD, P. R. 2012. Exome sequencing and genetic testing for MODY. *PLoS One,* 7**,** e38050.

JOHNSON, J. O., MANDRIOLI, J., BENATAR, M., ABRAMZON, Y., VAN DEERLIN, V. M., TROJANOWSKI, J. Q., GIBBS, J. R., BRUNETTI, M., GRONKA, S., WUU, J., DING, J., MCCLUSKEY, L., MARTINEZ-LAGE, M., FALCONE, D., HERNANDEZ, D. G., AREPALLI, S., CHONG, S., SCHYMICK, J. C., ROTHSTEIN, J., LANDI, F., WANG, Y. D., CALVO, A., MORA, G., SABATELLI, M., MONSURRO, M. R., BATTISTINI, S., SALVI, F., SPATARO, R., SOLA, P., BORGHERO, G., CONSORTIUM, I., GALASSI, G., SCHOLZ, S. W., TAYLOR, J. P., RESTAGNO, G., CHIO, A. & TRAYNOR, B. J. 2010. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron,* 68**,** 857-64.

JOHNSON, S. S., ZAIKOVA, E., GOERLITZ, D. S., BAI, Y. & TIGHE, S. W. 2017. Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer. *Journal of Biomolecular Techniques : JBT,* 28**,** 2-7.

JONES, H. D. 2015. Regulatory uncertainty over genome editing. *Nature Plants,* 1**,** 14011.

JOO, H. S., DONALDSON-WOOD, C. R. & JOHNSON, R. H. 1976. Observations on the pathogenesis of porcine parvovirus infection. *Arch Virol,* 51**,** 123-9.

JOSHI, N. & FASS, J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Available at https://github.com/najoshi/sickle.

JUNG, S.-N., LIM, H. S., LIU, L., CHANG, J. W., LIM, Y. C., RHA, K. S. & KOO, B. S. 2018. LAMB3 mediates metastatic tumor behavior in papillary thyroid cancer by regulating c-MET/Akt signals. *Scientific Reports,* 8**,** 2718.

KAPETANOVIC, R., FAIRBAIRN, L., BERALDI, D., SESTER, D. P., ARCHIBALD, A. L., TUGGLE, C. K. & HUME, D. A. 2012. Pig bone marrow-derived macrophages resemble human macrophages in their response to bacterial lipopolysaccharide. *J Immunol,* 188**,** 3382-94.

KEANE, T. M., GOODSTADT, L., DANECEK, P., WHITE, M. A., WONG, K., YALCIN, B., HEGER, A., AGAM, A., SLATER, G., GOODSON, M., FURLOTTE, N. A., ESKIN, E., NELLAKER, C., WHITLEY, H., CLEAK, J., JANOWITZ, D., HERNANDEZ-PLIEGO, P., EDWARDS, A., BELGARD, T. G., OLIVER, P. L., MCINTYRE, R. E., BHOMRA, A.,

NICOD, J., GAN, X., YUAN, W., VAN DER WEYDEN, L., STEWARD, C. A., BALA, S., STALKER, J., MOTT, R., DURBIN, R., JACKSON, I. J., CZECHANSKI, A., GUERRA-ASSUNCAO, J. A., DONAHUE, L. R., REINHOLDT, L. G., PAYSEUR, B. A., PONTING, C. P., BIRNEY, E., FLINT, J. & ADAMS, D. J. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature,* 477**,** 289-294.

KEEL, B. N., NONNEMAN, D. J. & ROHRER, G. A. 2017. A survey of single nucleotide polymorphisms identified from whole-genome sequencing and their functional effect in the porcine genome. *Animal Genetics,* 48**,** 404-411.

KIISKI, J. I., PELTTARI, L. M., KHAN, S., FREYSTEINSDOTTIR, E. S., REYNISDOTTIR, I., HART, S. N., SHIMELIS, H., VILSKE, S., KALLIONIEMI, A., SCHLEUTKER, J., LEMINEN, A., BÜTZOW, R., BLOMQVIST, C., BARKARDOTTIR, R. B., COUCH, F. J., AITTOMÄKI, K. & NEVANLINNA, H. 2014. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 15172-15177.

KIM, J., LARKIN, D. M., CAI, Q., ASAN, ZHANG, Y., GE, R.-L., AUVIL, L., CAPITANU, B., ZHANG, G., LEWIN, H. A. & MA, J. 2013. Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences,* 110**,** 1785-1790.

KITZMAN, J. O. 2016. Haplotypes drop by drop. *Nat Biotech,* 34**,** 296-298.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D. & LIN, L. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22.

KOLMOGOROV, M., RANEY, B., PATEN, B. & PHAM, S. 2014. Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics,* 30**,** i302-i309.

KOLTE, A. M., NIELSEN, H. S., MOLTKE, I., DEGN, B., PEDERSEN, B., SUNDE, L., NIELSEN, F. C. & CHRISTIANSEN, O. B. 2011. A genome-wide scan in affected sibling pairs with idiopathic recurrent miscarriage suggests genetic linkage. *Mol Hum Reprod,* 17**,** 379-85.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R. & STEFANSSON, K. 2002. A high-resolution recombination map of the human genome. *Nat Genet,* 31**,** 241-7.

KOREN, S., RHIE, A., WALENZ, B. P., DILTHEY, A. T., BICKHART, D. M., KINGAN, S. B., HIENDLEDER, S., WILLIAMS, J. L., SMITH, T. P. L. & PHILLIPPY, A. 2018. Complete assembly of parental haplotypes with trio binning. *bioRxiv*.

KOREN, S., SCHATZ, M. C., WALENZ, B. P., MARTIN, J., HOWARD, J. T. & GANAPATHY, G. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol,* 30.

KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H. & PHILLIPPY, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*.

KORLACH, J. 2015. *Understanding accuracy in SMRT sequencing* [Online]. Available: http://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf [Accessed 08/06/2018].

KRASNOV, K., V., TZETIS, M., CHENG, J., GUGGINO, W., B. & CUTTING, G., R. 2008. Localization studies of rare missense mutations in cystic fibrosis transmembrane conductance regulator (CFTR) facilitate interpretation of genotype-phenotype relationships. *Human Mutation,* 29**,** 1364-1372.

KRAWCZAK, M., REISS, J. & COOPER, D. N. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet,* 90**,** 41-54.

KRISTENSEN, L., STOIER, S., WURTZ, J. & HINRICHSEN, L. 2014. Trends in meat science and technology: the future looks bright, but the journey will be long. *Meat Sci,* 98**,** 322-9.

KRONENBERG, Z. N., HALL, R. J., HIENDLEDER, S., SMITH, T. P. L., SULLIVAN, S. T., WILLIAMS, J. L. & KINGAN, S. B. 2018. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*.

KUMAR, P., HENIKOFF, S. & NG, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols,* 4**,** 1073-1081.

KUMAR, V., KIM, K., JOSEPH, C., KOURRICH, S., YOO, S.-H., HUANG, H. C., VITATERNA, M. H., PARDO-MANUEL DE VILLENA, F., CHURCHILL, G., BONCI, A. & TAKAHASHI, J. S. 2013. C57BL/6N Mutation in Cytoplasmic FMRP interacting protein 2 Regulates Cocaine Response. *Science,* 342**,** 1508-1512.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biology,* 5**,** 1-9.

LACHANCE, J. 2009. Detecting selection-induced departures from Hardy-Weinberg proportions. *Genetics, Selection, Evolution : GSE,* 41**,** 15-15.

LAERE, A. S., NGUYEN, M., BRAUNSCHWEIG, M., NEZER, C., COLLETTE, C. & MOREAU, L. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature,* 425.

LAIRD, N. M. & LANGE, C. 2011. Principles of Inheritance: Mendel's Laws and Genetic Models. *In:* LAIRD, N. M. & LANGE, C. (eds.) *The Fundamentals of Modern Statistical Genetics.* New York, NY: Springer New York.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LARSEN, E., GRAN, C., SÆTHER, B. E., SEEBERG, E. & KLUNGLAND, A. 2003. Proliferation Failure and Gamma Radiation Sensitivity of Fen1 Null Mutant Mice at the Blastocyst Stage. *Molecular and Cellular Biology,* 23**,** 5346-5353.

LAYER, R. M., CHIANG, C., QUINLAN, A. R. & HALL, I. M. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology,* 15**,** R84.

LEBLOND, J., LAPRISE, M.-H., GAUDREAU, S., GRONDIN, F., KISIEL, W. & DUBOIS, C. M. 2006. The serpin proteinase inhibitor 8: An endogenous furin inhibitor released from human platelets. *Thromb Haemost,* 96**,** 243-252.

LEBRUN, J.-J. 2012. The Dual Role of TGF in Human Cancer: From Tumor Suppression to Cancer Metastasis. *ISRN Molecular Biology,* 2012**,** 28.

LEE, C. Y., QIN, J., MUNYARD, K. A., SIVA SUBRAMANIAM, N., WETHERALL, J. D., STEAR, M. J. & GROTH, D. M. 2012. Conserved haplotype blocks within the sheep MHC and low SNP heterozygosity in the Class IIa subregion. *Animal Genetics,* 43**,** 429-437.

LEE, H., DEIGNAN, J. L., DORRANI, N., STROM, S. P., KANTARCI, S., QUINTERO-RIVERA, F., DAS, K., TOY, T., HARRY, B., YOURSHAW, M., FOX, M., FOGEL, B. L., MARTINEZ-AGOSTO, J. A., WONG, D. A., CHANG, V. Y., SHIEH, P. B., PALMER, C. G., DIPPLE, K. M., GRODY, W. W., VILAIN, E. & NELSON, S. F. 2014. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA,* 312**,** 1880-7.

LEFEBVRE, R. C. 2015. Fetal mummification in the major domestic species: current perspectives on causes and management. *Veterinary Medicine: Research and Reports,* 2105:6**,** 233-244.

LEITCH, I. J., FAY, M. F. & PELLICER, J. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society,* 164**,** 10-15.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H.-H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T.,

WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION, C. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285.
LELIEVELD, S. H., SPIELMANN, M., MUNDLOS, S., VELTMAN, J. A. & GILISSEN, C. 2015. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation,* 36**,** 815-822.
LENNON, M. J., JONES, S. P., LOVELACE, M. D., GUILLEMIN, G. J. & BREW, B. J. 2017. Bcl11b—A Critical Neurodevelopmental Transcription Factor—Roles in Health and Disease. *Frontiers in Cellular Neuroscience,* 11.
LESCAI, F., MARASCO, E., BACCHELLI, C., STANIER, P., MANTOVANI, V. & BEALES, P. 2014. Identification and validation of loss of function variants in clinical contexts. *Molecular Genetics & Genomic Medicine,* 2**,** 58-63.
LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-93.
LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997v1 [q-bio.GN]*.
LI, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics,* 32**,** 2103-2110.
LI, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics***,** bty191-bty191.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J. & HOMER, N. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25.
LI, H. & HOMER, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform,* 11**,** 473-83.
LI, H., RUAN, J. & DURBIN, R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res,* 18**,** 1851-8.
LI, M., CHEN, L., TIAN, S., LIN, Y., TANG, Q., ZHOU, X., LI, D., YEUNG, C. K. L., CHE, T., JIN, L., FU, Y., MA, J., WANG, X., JIANG, A., LAN, J., PAN, Q., LIU, Y., LUO, Z., GUO, Z., LIU, H., ZHU, L., SHUAI, S., TANG, G., ZHAO, J., JIANG, Y., BAI, L., ZHANG, S., MAI, M., LI, C., WANG, D., GU, Y., WANG, G., LU, H., LI, Y., ZHU, H., LI, Z., LI, M., GLADYSHEV, V. N., JIANG, Z., ZHAO, S., WANG, J., LI, R. & LI, X. 2017a. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Research,* 27**,** 865-874.
LI, M., TIAN, S., JIN, L., ZHOU, G., LI, Y. & ZHANG, Y. 2013a. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet,* 45.

LI, R., LI, Y., FANG, X., YANG, H., WANG, J., KRISTIANSEN, K. & WANG, J. 2009b. SNP detection for massively parallel whole-genome resequencing. *Genome Research,* 19**,** 1124-1132.
LI, R., LI, Y., KRISTIANSEN, K. & WANG, J. 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics,* 24**,** 713-714.
LI, R., YU, C., LI, Y., LAM, T.-W., YIU, S.-M., KRISTIANSEN, K. & WANG, J. 2009c. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-1967.
LI, S., LI, R., LI, H., LU, J., LI, Y., BOLUND, L., SCHIERUP, M. H. & WANG, J. 2013b. SOAPindel: Efficient identification of indels from short paired reads. *Genome Research,* 23**,** 195-200.
LI, S., O'NEILL, S. R. S., ZHANG, Y., HOLTZMAN, M. J., TAKEMARU, K.-I., KORACH, K. S. & WINUTHAYANON, W. 2017b. Estrogen receptor α is required for oviductal transport of embryos. *The FASEB Journal,* 31**,** 1595-1607.
LI, W., HUANG, Q., SUN, D., ZHANG, G. & TAN, J. 2017c. RDM1 gene overexpression represents a therapeutic target in papillary thyroid carcinoma. *Endocr Connect,* 6**,** 700-707.
LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X., LIU, B., YANG, B. & FAN, W. 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics,* 11**,** 25-37.
LI, Z., GOU, J., JIA, J. & ZHAO, X. 2015. MicroRNA-429 functions as a regulator of epithelial–mesenchymal transition by targeting Pcdh8 during murine embryo implantation. *Human Reproduction,* 30**,** 507-518.
LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., SANDSTROM, R., BERNSTEIN, B., BENDER, M. A., GROUDINE, M., GNIRKE, A., STAMATOYANNOPOULOS, J., MIRNY, L. A., LANDER, E. S. & DEKKER, J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* 326.
LIM, J.-Q., TENNAKOON, C., GUAN, P. & SUNG, W.-K. 2015. BatAlign: an incremental method for accurate alignment of sequencing reads. *Nucleic Acids Research,* 43**,** e107-e107.
LIM, K. H. & FAIRBROTHER, W. G. 2012. Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics,* 28**,** 1031-1032.
LINDBLAD-TOH, K., WADE, C. M., MIKKELSEN, T. S., KARLSSON, E. K., JAFFE, D. B., KAMAL, M., CLAMP, M., CHANG, J. L., KULBOKAS, E. J. R. D., ZODY, M. C., MAUCELI, E., XIE, X., BREEN, M., WAYNE, R. K., OSTRANDER, E. A., PONTING, C. P., GALIBERT, F., SMITH, D. R., DEJONG, P. J., KIRKNESS, E., ALVAREZ, P., BIAGI, T., BROCKMAN, W., BUTLER, J., CHIN, C. W., COOK, A., CUFF, J., DALY, M. J., DECAPRIO, D. & GNERRE, S. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature,* 438.

LIU, Q., SHEN, E., MIN, Q., LI, X., WANG, X., LI, X., SUN, Z. S. & WU, J. 2012. Exome-assistant: a rapid and easy detection of disease-related genes and genetic variations from exome sequencing. *BMC Genomics,* 13**,** 692.
LIU, Y., KOYUTÜRK, M., MAXWELL, S., XIANG, M., VEIGL, M., COOPER, R. S., TAYO, B. O., LI, L., LAFRAMBOISE, T., WANG, Z., ZHU, X. & CHANCE, M. R. 2014. Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. *BMC Genomics,* 15**,** 685.
LOMAN, N. J., QUICK, J. & SIMPSON, J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Meth,* 12**,** 733-735.
LOMAN, N. J. & WATSON, M. 2015. Successful test launch for nanopore sequencing. *Nature Methods,* 12**,** 303.
LÓPEZ, C. & CARMONA, J. U. 2010. Uterine torsion diagnosed in a mare at 515 days' gestation. *Equine Veterinary Education,* 22**,** 483-486.
LOVE, R., KIRKLAND, P., MORILLA, A., YOON, K.-J. & ZIMMERMAN, J. J. 2008. Menangle Virus: A New Cause of Fetal Mummification and Congenital Defects in Pigs. *Trends in Emerging Viral Infections of Swine.* Iowa State Press.
LU, H., GIORDANO, F. & NING, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics,* 14**,** 265-279.
LU, L., HOU, X., SHI, S., KÖRNER, C. & STANLEY, P. 2010. Slc35c2 Promotes Notch1 Fucosylation and Is Required for Optimal Notch Signaling in Mammalian Cells. *The Journal of Biological Chemistry,* 285**,** 36245-36254.
LU, Y. C., CHEN, H., FOK, K. L., TSANG, L. L., YU, M. K., ZHANG, X. H., CHEN, J., JIANG, X., CHUNG, Y. W., MA, A. C. H., LEUNG, A. Y. H., HUANG, H. F. & CHAN, H. C. 2012. CFTR mediates bicarbonate-dependent activation of miR-125b in preimplantation embryo development. *Cell Research,* 22**,** 1453.
LUDWIG, D. L., MACINNES, M. A., TAKIGUCHI, Y., PURTYMUN, P. E., HENRIE, M., FLANNERY, M., MENESES, J., PEDERSEN, R. A. & CHEN, D. J. 1998. A murine AP-endonuclease gene-targeted deficiency with post-implantation embryonic progression and ionizing radiation sensitivity. *Mutation Research/DNA Repair,* 409**,** 17-29.
LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y., TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, D. W., YIU, S.-M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T.-W. & WANG, J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1**,** 18.
MA, J., LI, M., WANG, H. & LI, X. 2012. Genotyping of the porcine ryanodine receptor 1 (RYR1) and estrogen receptor 1 (ESR1) genes by high resolution melting (HRM) approach. *Biotechnology and Bioprocess Engineering,* 17**,** 1076-1079.

MACARTHUR, D. G., BALASUBRAMANIAN, S., FRANKISH, A., HUANG, N., MORRIS, J., WALTER, K., JOSTINS, L., HABEGGER, L., PICKRELL, J. K., MONTGOMERY, S. B., ALBERS, C. A., ZHANG, Z. D., CONRAD, D. F., LUNTER, G., ZHENG, H., AYUB, Q., DEPRISTO, M. A., BANKS, E., HU, M., HANDSAKER, R. E., ROSENFELD, J. A., FROMER, M., JIN, M., MU, X. J., KHURANA, E., YE, K., KAY, M., SAUNDERS, G. I., SUNER, M.-M., HUNT, T., BARNES, I. H. A., AMID, C., CARVALHO-SILVA, D. R., BIGNELL, A. H., SNOW, C., YNGVADOTTIR, B., BUMPSTEAD, S., COOPER, D. N., XUE, Y., ROMERO, I. G., CONSORTIUM, G. P., WANG, J., LI, Y., GIBBS, R. A., MCCARROLL, S. A., DERMITZAKIS, E. T., PRITCHARD, J. K., BARRETT, J. C., HARROW, J., HURLES, M. E., GERSTEIN, M. B. & TYLER-SMITH, C. 2012. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science,* 335**,** 823-828.

MACARTHUR, D. G. & TYLER-SMITH, C. 2010. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet,* 19**,** R125-30.

MACFARLANE, A. J., PERRY, C. A., GIRNARY, H. H., GAO, D., ALLEN, R. H., STABLER, S. P., SHANE, B. & STOVER, P. J. 2009. Mthfd1 Is an Essential Gene in Mice and Alters Biomarkers of Impaired One-carbon Metabolism. *Journal of Biological Chemistry,* 284**,** 1533-1539.

MAJEWSKI, J., SCHWARTZENTRUBER, J., LALONDE, E., MONTPETIT, A. & JABADO, N. 2011. What can exome sequencing do for you? *J Med Genet,* 48**,** 580-9.

MAKINO, T., RUBIN, C.-J., CARNEIRO, M., AXELSSON, E., ANDERSSON, L. & WEBSTER, M. T. 2018. Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants. *Genome Biology and Evolution,* 10**,** 276-290.

MALINDA, K. M. & KLEINMAN, H. K. 1996. The laminins. *Int J Biochem Cell Biol,* 28**,** 957-9.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics,* 39**,** 906.

MARDIS, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics,* 24**,** 133-141.

MARDIS, E. R. 2010. The $1,000 genome, the $100,000 analysis? *Genome Medicine,* 2**,** 84-84.

MARQUES, M. D., CRITCHLEY, C. R. & WALSHE, J. 2014. Attitudes to genetically modified food over time: How trust in organizations and the media cycle predict support. *Public Understanding of Science,* 24**,** 601-618.

MARSCHALL, T., COSTA, I. G., CANZAR, S., BAUER, M., KLAU, G. W., SCHLIEP, A. & SCHÖNHUTH, A. 2012. CLEVER: clique-enumerating variant finder. *Bioinformatics,* 28**,** 2875-2882.

MARSTON, D. A., MCELHINNEY, L. M., ELLIS, R. J., HORTON, D. L., WISE, E. L., LEECH, S. L., DAVID, D., DE LAMBALLERIE, X. & FOOKS, A. R. 2013. Next generation sequencing of viral RNA genomes. *BMC Genomics,* 14**,** 444-444.

MASCHER, M., GUNDLACH, H., HIMMELBACH, A., BEIER, S., TWARDZIOK, S. O., WICKER, T., RADCHUK, V., DOCKTER, C., HEDLEY, P. E., RUSSELL, J., BAYER, M., RAMSAY, L., LIU, H., HABERER, G., ZHANG, X.-Q., ZHANG, Q., BARRERO, R. A., LI, L., TAUDIEN, S., GROTH, M., FELDER, M., HASTIE, A., ŠIMKOVÁ, H., STAŇKOVÁ, H., VRÁNA, J., CHAN, S., MUÑOZ-AMATRIAÍN, M., OUNIT, R., WANAMAKER, S., BOLSER, D., COLMSEE, C., SCHMUTZER, T., ALIYEVA-SCHNORR, L., GRASSO, S., TANSKANEN, J., CHAILYAN, A., SAMPATH, D., HEAVENS, D., CLISSOLD, L., CAO, S., CHAPMAN, B., DAI, F., HAN, Y., LI, H., LI, X., LIN, C., MCCOOKE, J. K., TAN, C., WANG, P., WANG, S., YIN, S., ZHOU, G., POLAND, J. A., BELLGARD, M. I., BORISJUK, L., HOUBEN, A., DOLEŽEL, J., AYLING, S., LONARDI, S., KERSEY, P., LANGRIDGE, P., MUEHLBAUER, G. J., CLARK, M. D., CACCAMO, M., SCHULMAN, A. H., MAYER, K. F. X., PLATZER, M., CLOSE, T. J., SCHOLZ, U., HANSSON, M., ZHANG, G., BRAUMANN, I., SPANNAGL, M., LI, C., WAUGH, R. & STEIN, N. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature,* 544**,** 427-433.

MASCHER, M., JOST, M., KUON, J.-E., HIMMELBACH, A., AßFALG, A., BEIER, S., SCHOLZ, U., GRANER, A. & STEIN, N. 2014. Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biology,* 15**,** R78-R78.

MASCHER, M., RICHMOND, T. A., GERHARDT, D. J., HIMMELBACH, A., CLISSOLD, L., SAMPATH, D., AYLING, S., STEUERNAGEL, B., PFEIFER, M., D'ASCENZO, M., AKHUNOV, E. D., HEDLEY, P. E., GONZALES, A. M., MORRELL, P. L., KILIAN, B., BLATTNER, F. R., SCHOLZ, U., MAYER, K. F., FLAVELL, A. J., MUEHLBAUER, G. J., WAUGH, R., JEDDELOH, J. A. & STEIN, N. 2013. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *Plant J,* 76**,** 494-505.

MATHEW, D. J., LUCY, M. C. & D GEISERT, R. 2016. Interleukins, interferons, and establishment of pregnancy in pigs. *Reproduction,* 151**,** R111-R122.

MATUKUMALLI, L. K., LAWLEY, C. T., SCHNABEL, R. D., TAYLOR, J. F., ALLAN, M. F., HEATON, M. P., O'CONNELL, J., MOORE, S. S., SMITH, T. P. L., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLOS ONE,* 4**,** e5350.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H. Z., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E., DIEGEL, M., BATES, D., HANSEN, R. S., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. Systematic Localization of

Common Disease-Associated Variation in Regulatory DNA. *Science,* 337**,** 1190-1195.
MAXAM, A. M. & GILBERT, W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 560-564.
MCCALLIE, B. R., PARKS, J. C., PATTON, A. L., GRIFFIN, D. K., SCHOOLCRAFT, W. B. & KATZ-JAFFE, M. G. 2016. Hypomethylation and Genetic Instability in Monosomy Blastocysts May Contribute to Decreased Implantation Potential. *PLOS ONE,* 11**,** e0159507.
MCCARTHY, D. J., HUMBURG, P., KANAPIN, A., RIVAS, M. A., GAULTON, K., CAZIER, J. B., DONNELLY, P. & CONSORTIUM, W. G. S. 2014. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine,* 6**,** 15.
MCCLURE, M. C., BICKHART, D., NULL, D., VANRADEN, P., XU, L., WIGGANS, G., LIU, G., SCHROEDER, S., GLASSCOCK, J., ARMSTRONG, J., COLE, J. B., VAN TASSELL, C. P. & SONSTEGARD, T. S. 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. *PLoS One,* 9**,** e92769.
MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S. & DALY, M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20.
MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R. S., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biology,* 17**,** 122.
MCLAREN, W., PRITCHARD, B., RIOS, D., CHEN, Y., FLICEK, P. & CUNNINGHAM, F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-70.
MCMAHON, S. B., VAN BUSKIRK, H. A., DUGAN, K. A., COPELAND, T. D. & COLE, M. D. 1998. The Novel ATM-Related Protein TRRAP Is an Essential Cofactor for the c-Myc and E2F Oncoproteins. *Cell,* 94**,** 363-374.
MEMON, M. A., ANWAY, M. D., COVERT, T. R., UZUMCU, M. & SKINNER, M. K. 2008. Transforming growth factor beta (TGFβ1, TGFβ2 and TGFβ3) null-mutant phenotypes in embryonic gonadal development. *Molecular and Cellular Endocrinology,* 294**,** 70-80.
MENASHA, J., LEVY, B., HIRSCHHORN, K. & KARDON, N. B. 2005. Incidence and spectrum of chromosome abnormalities in spontaneous abortions: new insights from a 12-year study. *Genet Med,* 7**,** 251-63.
MENG, L., PAMMI, M., SARONWALA, A. & ET AL. 2017a. Use of exome sequencing for infants in intensive care units: Ascertainment of severe single-gene disorders and effect on medical management. *JAMA Pediatrics***,** e173438.

MENG, Q., WANG, K., LIU, X., ZHOU, H., XU, L., WANG, Z. & FANG, M. 2017b. Identification of growth trait related genes in a Yorkshire purebred pig population by genome wide association studies. *Asian-Australas J Anim Sci,* 30**,** 462-469.
MENGELING, W. L., CUTLIP, R. C., WILSON, R. A., PARKS, J. B. & MARSHALL, R. F. 1975. Fetal mummification associated with porcine parvovirus infection. *J Am Vet Med Assoc,* 166**,** 993-5.
MERKER, J. D., WENGER, A. M., SNEDDON, T., GROVE, M., ZAPPALA, Z., FRESARD, L., WAGGOTT, D., UTIRAMERUR, S., HOU, Y., SMITH, K. S., MONTGOMERY, S. B., WHEELER, M., BUCHAN, J. G., LAMBERT, C. C., ENG, K. S., HICKEY, L., KORLACH, J., FORD, J. & ASHLEY, E. A. 2017. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics In Medicine,* 20**,** 159.
MEUNIER, J. & DURET, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol,* 21**,** 984-90.
MEURENS, F., SUMMERFIELD, A., NAUWYNCK, H., SAIF, L. & GERDTS, V. 2012. The pig: a model for human infectious diseases. *Trends Microbiol,* 20**,** 50-7.
MEUWISSEN, T. H., HAYES, B. J. & GODDARD, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157**,** 1819-1829.
MEUWISSEN, T. H. E. & GODDARD, M. E. 1996. The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution : GSE,* 28**,** 161-176.
MEYNERT, A. M., ANSARI, M., FITZPATRICK, D. R. & TAYLOR, M. S. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics,* 15**,** 247.
MILLER, C. A., HAMPTON, O., COARFA, C. & MILOSAVLJEVIC, A. 2011. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLOS ONE,* 6**,** e16327.
MISRA, M. K., MISHRA, A., PHADKE, S. R. & AGRAWAL, S. 2016. Association of functional genetic variants of CTLA4 with reduced serum CTLA4 protein levels and increased risk of idiopathic recurrent miscarriages. *Fertil Steril,* 106**,** 1115-1123.e6.
MISTRY, V., BOCKETT, N. A., LEVINE, A. P., MIRZA, M. M., HUNT, K. A., CICLITIRA, P. J., HUMMERICH, H., NEUHAUSEN, S. L., SIMPSON, M. A., PLAGNOL, V. & VAN HEEL, D. A. 2015. Exome Sequencing of 75 Individuals from Multiply Affected Coeliac Families and Large Scale Resequencing Follow Up. *PLoS ONE,* 10**,** e0116845.
MONCUNILL, V., GONZALEZ, S., BEA, S., ANDRIEUX, L. O., SALAVERRIA, I., ROYO, C., MARTINEZ, L., PUIGGROS, M., SEGURA-WANG, M., STUTZ, A. M., NAVARRO, A., ROYO, R., GELPI, J. L., GUT, I. G., LOPEZ-OTIN, C., OROZCO, M., KORBEL, J. O., CAMPO, E., PUENTE, X. S. & TORRENTS, D. 2014. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotech,* 32**,** 1106-1112.

MOORE, A. A. & RICHARDSON, G. F. 1995. Uterine torsion and fetal mummification in a cow. *The Canadian Veterinary Journal,* 36**,** 705-706.
MORRELL, P. L., TOLENO, D. M., LUNDY, K. E. & CLEGG, M. T. 2006. Estimating the Contribution of Mutation, Recombination and Gene Conversion in the Generation of Haplotypic Diversity. *Genetics,* 173**,** 1705-1723.
MOSER, C., LANG, S. A., HACKL, C., ZHANG, H., LUNDGREN, K., HONG, V., MCKENZIE, A., WEBER, B., PARK, J. S., SCHLITT, H. J., GEISSLER, E. K., JUNG, Y. D. & STOELTZING, O. 2012. Oncogenic MST1R Activity in Pancreatic and Gastric Cancer Represents a Valid Target of HSP90 Inhibitors. *Anticancer Research,* 32**,** 427-437.
MU, X. J., LU, Z. J., KONG, Y., LAM, H. Y. K. & GERSTEIN, M. B. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Research,* 39**,** 7058-7076.
MUIR, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics,* 124**,** 342-355.
MUNOZ, G., OVILO, C., ESTELLE, J., SILIO, L., FERNANDEZ, A. & RODRIGUEZ, C. 2007. Association with litter size of new polymorphisms on ESR1 and ESR2 genes in a Chinese-European pig line. *Genet Sel Evol,* 39**,** 195-206.
MURAMATSU, Y., LEJUKOLE, H. Y., TANIGUCHI, Y., KONFORTOV, B. A., YAMADA, T., YASUE, H. & SASAKI, Y. 2002. Chromosomal assignments of expressed sequence tags for ACTG1, AHSG, COL1A1, GNAS1, and RPLP1 expressed abundantly in the bovine foetus. *Anim Genet,* 33**,** 230-1.
MURAOKA, R. S., SUN, W. Y., COLBERT, M. C., WALTZ, S. E., WITTE, D. P., DEGEN, J. L. & DEGEN, S. J. F. 1999. The Ron/STK receptor tyrosine kinase is essential for peri-implantation development in the mouse. *The Journal of Clinical Investigation,* 103**,** 1277-1285.
MUTARELLI, M., MARWAH, V., RISPOLI, R., CARRELLA, D., DHARMALINGAM, G., OLIVA, G. & DI BERNARDO, D. 2014. A community-based resource for automatic exome variant-calling and annotation in Mendelian disorders. *BMC Genomics,* 15 Suppl 3**,** S5.
MYERS, E. W. 2005. The fragment assembly string graph. *Bioinformatics,* 21**,** ii79-ii85.
MYERS, E. W., SUTTON, G. G., DELCHER, A. L., DEW, I. M., FASULO, D. P. & FLANIGAN, M. J. 2000. A whole-genome assembly of Drosophila. *Science,* 287.
MYERS, S., FREEMAN, C., AUTON, A., DONNELLY, P. & MCVEAN, G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics,* 40**,** 1124.
NADALIN, F., VEZZI, F. & POLICRITI, A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics,* 13**,** S8.

NAGARAJAN, N. & POP, M. 2013. Sequence assembly demystified. *Nat Rev Genet,* 14**,** 157-167.
NAKAYAMA, T., SOMA, M., WATANABE, Y., HASIMU, B., SATO, M., AOI, N., KOSUGE, K., KANMATSUSE, K., KOKUBUN, S., MARROW, J. D. & OATES, J. A. 2002. Splicing mutation of the prostacyclin synthase gene in a family associated with hypertension. *Biochemical and Biophysical Research Communications,* 297**,** 1135-1139.
NALLA, V. K. & ROGAN, P. K. 2005. Automated splicing mutation analysis by information theory. *Human Mutation,* 25**,** 334-342.
NATIONAL RESEARCH COUNCIL 1988. *Mapping and Sequencing the Human Genome,* Washington, DC, The National Academies Press.
NG, P. C. & HENIKOFF, S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research,* 31**,** 3812-3814.
NGUYEN, D. T., LEE, K., CHOI, H., CHOI, M.-K., LE, M. T., SONG, N., KIM, J.-H., SEO, H. G., OH, J.-W., LEE, K., KIM, T.-H. & PARK, C. 2012. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. *BMC Genomics,* 13**,** 584-584.
NIELSEN, D. M., EHM, M. G. & WEIR, B. S. 1998. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet,* 63**,** 1531-40.
NIU, D., WEI, H.-J., LIN, L., GEORGE, H., WANG, T., LEE, I. H., ZHAO, H.-Y., WANG, Y., KAN, Y., SHROCK, E., LESHA, E., WANG, G., LUO, Y., QING, Y., JIAO, D., ZHAO, H., ZHOU, X., WANG, S., WEI, H., GÜELL, M., CHURCH, G. M. & YANG, L. 2017. Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science,* 357**,** 1303.
NOLL, A. C., MILLER, N. A., SMITH, L. D., YOO, B., FIEDLER, S., COOLEY, L. D., WILLIG, L. K., PETRIKIN, J. E., CAKICI, J., LESKO, J., NEWTON, A., DETHERAGE, K., THIFFAULT, I., SAUNDERS, C. J., FARROW, E. G. & KINGSMORE, S. F. 2016. Clinical detection of deletion structural variants in whole-genome sequences. *Npj Genomic Medicine,* 1**,** 16026.
NOONAN, J. P. 2010. Neanderthal genomics and the evolution of modern humans. *Genome Research,* 20**,** 547-553.
NSENGIMANA, J., BARET, P., HALEY, C. S. & VISSCHER, P. M. 2004. Linkage Disequilibrium in the Domesticated Pig. *Genetics,* 166**,** 1395-1404.
NYMAN, L. R., COX, K. B., HOPPEL, C. L., KERNER, J., BARNOSKI, B. L., HAMM, D. A., TIAN, L., SCHOEB, T. R. & WOOD, P. A. 2005. Homozygous carnitine palmitoyltransferase 1a (liver isoform) deficiency is lethal in the mouse. *Molecular Genetics and Metabolism,* 86**,** 179-187.
O'RAWE, J., JIANG, T., SUN, G., WU, Y., WANG, W., HU, J., BODILY, P., TIAN, L., HAKONARSON, H., JOHNSON, W. E., WEI, Z., WANG, K. & LYON, G. J. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med,* 5**,** 28.

OBER, C., ALDRICH, C. L., CHERVONEVA, I., BILLSTRAND, C., RAHIMOV, F., GRAY, H. L. & HYSLOP, T. 2003. Variation in the HLA-G promoter region influences miscarriage rates. *Am J Hum Genet,* 72**,** 1425-35.

ONTERU, S. K., FAN, B., DU, Z. Q., GARRICK, D. J., STALDER, K. J. & ROTHSCHILD, M. F. 2012. A whole-genome association study for pig reproductive traits. *Anim Genet,* 43**,** 18-26.

ORELIO, C., HAAK, E., PEETERS, M. & DZIERZAK, E. 2008. Interleukin-1–mediated hematopoietic cell regulation in the aorta-gonad-mesonephros region of the mouse embryo. *Blood,* 112**,** 4895.

OSOEGAWA, K., ZHU, B., SHU, C. L., REN, T., CAO, Q., VESSERE, G. M., LUTZ, M. M., JENSEN-SEAMAN, M. I., ZHAO, S. & DE JONG, P. J. 2004. BAC Resources for the Rat Genome Project. *Genome Research,* 14**,** 780-785.

OZ MAMMAL GENOMES. 2016. *Oz Mammal Genomes - Bioplatforms Australia* [Online]. Available: http://www.bioplatforms.com/oz-mammals [Accessed].

PABINGER, S., DANDER, A., FISCHER, M., SNAJDER, R., SPERK, M., EFREMOVA, M., KRABICHLER, B., SPEICHER, M. R., ZSCHOCKE, J. & TRAJANOSKI, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform,* 15**,** 256-78.

PAIGE, K. N. 2010. The Functional Genomics of Inbreeding Depression: A New Approach to an Old Problem. *BioScience,* 60**,** 267-277.

PAIGEN, K. & PETKOV, P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet,* 11**,** 221-33.

PANKIN, A., CAMPOLI, C., DONG, X., KILIAN, B., SHARMA, R., HIMMELBACH, A., SAINI, R., DAVIS, S. J., STEIN, N., SCHNEEBERGER, K. & VON KORFF, M. 2014. Mapping-by-Sequencing Identifies HvPHYTOCHROME C as a Candidate Gene for the early maturity 5 Locus Modulating the Circadian Clock and Photoperiodic Flowering in Barley. *Genetics,* 198**,** 383-396.

PARDI, G., MARCONI, A. M. & CETIN, I. 2002. Placental-fetal Interrelationship in IUGR Fetuses—A Review. *Placenta,* 23**,** S136-S141.

PARKS, M. & LAMBERT, D. 2015. Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics,* 16**,** 19.

PARLE-MCDERMOTT, A., PANGILINAN, F., MILLS, J. L., SIGNORE, C. C., MOLLOY, A. M., COTTER, A., CONLEY, M., COX, C., KIRKE, P. N., SCOTT, J. M. & BRODY, L. C. 2005. A polymorphism in the MTHFD1 gene increases a mother's risk of having an unexplained second trimester pregnancy loss. *MHR: Basic science of reproductive medicine,* 11**,** 477-480.

PARLE-MCDERMOTT, A., MILLS JAMES, L., KIRKE PEADAR, N., COX, C., SIGNORE CAROLINE, C., KIRKE, S., MOLLOY ANNE, M., O'LEARY VALERIE, B., PANGILINAN FAITH, J., O'HERLIHY, C., BRODY LAWRENCE, C. & SCOTT JOHN, M. 2005. MTHFD1 R653Q polymorphism is a maternal genetic risk factor for severe abruptio

placentae. *American Journal of Medical Genetics Part A,* 132A**,** 365-368.
PASZKIEWICZ, K. & STUDHOLME, D. J. 2010. De novo assembly of short sequence reads. *Briefings in Bioinformatics,* 11**,** 457-472.
PAUDEL, Y., MADSEN, O., MEGENS, H.-J., FRANTZ, L. A. F., BOSSE, M., BASTIAANSEN, J. W. M., CROOIJMANS, R. P. M. A. & GROENEN, M. A. M. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics,* 14**,** 449-449.
PAULINO, D., WARREN, R. L., VANDERVALK, B. P., RAYMOND, A., JACKMAN, S. D. & BIROL, I. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics,* 16**,** 230.
PAYNE, A., HOLMES, N., RAKYAN, V. & LOOSE, M. 2018. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*.
PEDERSEN, B. S., LAYER, R. M. & QUINLAN, A. R. 2016. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology,* 17**,** 118.
PENDLETON, M., SEBRA, R., PANG, A. W. C., UMMAT, A., FRANZEN, O., RAUSCH, T., STUTZ, A. M., STEDMAN, W., ANANTHARAMAN, T., HASTIE, A., DAI, H., FRITZ, M. H.-Y., CAO, H., COHAIN, A., DEIKUS, G., DURRETT, R. E., BLANCHARD, S. C., ALTMAN, R., CHIN, C.-S., GUO, Y., PAXINOS, E. E., KORBEL, J. O., DARNELL, R. B., MCCOMBIE, W. R., KWOK, P.-Y., MASON, C. E., SCHADT, E. E. & BASHIR, A. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Meth,* 12**,** 780-786.
PERLEBERG, C., KIND, A. & SCHNIEKE, A. 2018. Genetically engineered pigs as models for human disease. *Disease Models & Mechanisms,* 11.
PERTEA, M., LIN, X. & SALZBERG, S. L. 2001. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research,* 29**,** 1185-1190.
PERUCHO, L., ARTERO-CASTRO, A., GUERRERO, S., RAMÓN Y CAJAL, S., LLEONART, M. E. & WANG, Z.-Q. 2014. RPLP1, a Crucial Ribosomal Protein for Embryonic Development of the Nervous System. *PLOS ONE,* 9**,** e99956.
PEVZNER, P. A., TANG, H. & WATERMAN, M. S. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 9748-9753.
PHILIPP, T., PHILIPP, K., REINER, A., BEER, F. & KALOUSEK, D. K. 2003. Embryoscopic and cytogenetic analysis of 233 missed abortions: factors involved in the pathogenesis of developmental defects of early failed pregnancies. *Hum Reprod,* 18**,** 1724-32.
PHILLIPPY, A. M., SCHATZ, M. C. & POP, M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol,* 9**,** R55.
PIEL, F. B., ADAMKIEWICZ, T. V., AMENDAH, D., WILLIAMS, T. N., GUPTA, S. & GROSSE, S. D. 2016. Observed and expected frequencies of structural hemoglobin variants in newborn screening

surveys in Africa and the Middle East: Deviations from Hardy-Weinberg equilibrium. *Genetics in medicine : official journal of the American College of Medical Genetics,* 18**,** 265-274.

PIGORS, M., SARIG, O., HEINZ, L., PLAGNOL, V., FISCHER, J., MOHAMAD, J., MALCHIN, N., RAJPOPAT, S., KHARFI, M., LESTRINGANT, GILES G., SPRECHER, E., KELSELL, DAVID P. & BLAYDON, DIANA C. 2016. Loss-of-Function Mutations in SERPINB8 Linked to Exfoliative Ichthyosis with Impaired Mechanical Stability of Intercellular Adhesions. *American Journal of Human Genetics,* 99**,** 430-436.

PLANA, J., VAYREDA, M., VILARRASA, J., BASTONS, M., ROSELL, R., MARTINEZ, M., SAN GABRIEL, A., PUJOLS, J., BADIOLA, J. L., RAMOS, J. A. & DOMINGO, M. 1992. Porcine epidemic abortion and respiratory syndrome (mystery swine disease). Isolation in Spain of the causative agent and experimental reproduction of the disease. *Veterinary Microbiology,* 33**,** 203-211.

PONGPANICH, M., SULLIVAN, P. F. & TZENG, J.-Y. 2010. A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics,* 26**,** 1731-1737.

POP, M. 2009. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics,* 10**,** 354-366.

POPLIN, R., RUANO-RUBIO, V., DEPRISTO, M. A., FENNELL, T. J., CARNEIRO, M. O., VAN DER AUWERA, G. A., KLING, D. E., GAUTHIER, L. D., LEVY-MOONSHINE, A., ROAZEN, D., SHAKIR, K., THIBAULT, J., CHANDRAN, S., WHELAN, C., LEK, M., GABRIEL, S., DALY, M. J., NEALE, B., MACARTHUR, D. G. & BANKS, E. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

PORTO, A. G., BRUN, F., SEVERINI, G. M., LOSURDO, P., FABRIS, E., TAYLOR, M. R. G., MESTRONI, L. & SINAGRA, G. 2016. Clinical Spectrum of PRKAG2 Syndrome. *Circulation. Arrhythmia and electrophysiology,* 9**,** 10.1161/CIRCEP.115.003121 e003121.

PRATHER, R. S., LORSON, M., ROSS, J. W., WHYTE, J. J. & WALTERS, E. 2013. Genetically Engineered Pig Models for Human Diseases. *Annual Review of Animal Biosciences,* 1**,** 203-219.

PUCCI, S., ZONETTI, M. J., FISCO, T., POLIDORO, C., BOCCHINFUSO, G., PALLESCHI, A., NOVELLI, G., SPAGNOLI, L. G. & MAZZARELLI, P. 2016. Carnitine palmitoyl transferase-1A (CPT1A): a new tumor specific target in human breast cancer. *Oncotarget,* 7**,** 19982-96.

PUEBLA-OSORIO, N., LACEY, D. B., ALT, F. W. & ZHU, C. 2006. Early Embryonic Lethality Due to Targeted Inactivation of DNA Ligase III. *Molecular and Cellular Biology,* 26**,** 3935-3941.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, MANUEL A R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, PAUL I W., DALY, MARK J. & SHAM, PAK C. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics,* 81**,** 559-575.

PUTNAM, N. H., O'CONNELL, B. L., STITES, J. C., RICE, B. J., BLANCHETTE, M., CALEF, R., TROLL, C. J., FIELDS, A., HARTLEY, P. D., SUGNET, C. W., HAUSSLER, D., ROKHSAR, D. S. & GREEN, R. E. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*.

QUICK, J. 2018. *Ultra-long read sequencing protocol for RAD004* [Online]. protocols.io. Available: https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n [Accessed 16 May 2018].

QUICK, J., GRUBAUGH, N. D., PULLAN, S. T., CLARO, I. M., SMITH, A. D., GANGAVARAPU, K., OLIVEIRA, G., ROBLES-SIKISAKA, R., ROGERS, T. F., BEUTLER, N. A., BURTON, D. R., LEWIS-XIMENEZ, L. L., GOES DE JESUS, J., GIOVANETTI, M., HILL, S., BLACK, A., BEDFORD, T., CARROLL, M. W., NUNES, M., ALCANTARA, L. C., SABINO, E. C., BAYLIS, S. A., FARIA, N., LOOSE, M., SIMPSON, J. T., PYBUS, O. G., ANDERSEN, K. G. & LOMAN, N. J. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *bioRxiv*.

QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G., MIKHAIL, A., OUÉDRAOGO, N., AFROUGH, B., BAH, A., BAUM, J. H. J., BECKER-ZIAJA, B., BOETTCHER, J. P., CABEZA-CABRERIZO, M., CAMINO-SÁNCHEZ, Á., CARTER, L. L., DOERRBECKER, J., ENKIRCH, T., DORIVAL, I. G., HETZELT, N., HINZMANN, J., HOLM, T., KAFETZOPOULOU, L. E., KOROPOGUI, M., KOSGEY, A., KUISMA, E., LOGUE, C. H., MAZZARELLI, A., MEISEL, S., MERTENS, M., MICHEL, J., NGABO, D., NITZSCHE, K., PALLASCH, E., PATRONO, L. V., PORTMANN, J., REPITS, J. G., RICKETT, N. Y., SACHSE, A., SINGETHAN, K., VITORIANO, I., YEMANABERHAN, R. L., ZEKENG, E. G., RACINE, T., BELLO, A., SALL, A. A., FAYE, O., FAYE, O., MAGASSOUBA, N. F., WILLIAMS, C. V., AMBURGEY, V., WINONA, L., DAVIS, E., GERLACH, J., WASHINGTON, F., MONTEIL, V., JOURDAIN, M., BERERD, M., CAMARA, A., SOMLARE, H., CAMARA, A., GERARD, M., BADO, G., BAILLET, B., DELAUNE, D., NEBIE, K. Y., DIARRA, A., SAVANE, Y., PALLAWO, R. B., GUTIERREZ, G. J., MILHANO, N., ROGER, I., WILLIAMS, C. J., YATTARA, F., LEWANDOWSKI, K., TAYLOR, J., RACHWAL, P., J. TURNER, D., POLLAKIS, G., HISCOX, J. A., MATTHEWS, D. A., SHEA, M. K. O., JOHNSTON, A. M., WILSON, D., HUTLEY, E., SMIT, E., DI CARO, A., WÖLFEL, R., STOECKER, K., FLEISCHMANN, E., GABRIEL, M., WELLER, S. A., KOIVOGUI, L., DIALLO, B., KEÏTA, S., RAMBAUT, A., FORMENTY, P., et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature,* 530**,** 228-232.

QUINLAN, A. R. 2014. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr Protoc Bioinformatics.,* 47.

R DEVELOPMENT CORE TEAM. 2009. *R: A Language and Environment for Statistical Computing* [Online]. Vienna, Austria : the R Foundation for Statistical Computing. Available: http://www.R-project.org/ [Accessed].

RABBANI, B., MAHDIEH, N., HOSOMICHI, K., NAKAOKA, H. & INOUE, I. 2012. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet,* 57**,** 621-32.
RAGOUSSIS, J. 2009. Genotyping Technologies for Genetic Research. *Annual Review of Genomics and Human Genetics,* 10**,** 117-133.
RAMOS, A. M., CROOIJMANS, R. P. M. A., AFFARA, N. A., AMARAL, A. J., ARCHIBALD, A. L., BEEVER, J. E., BENDIXEN, C., CHURCHER, C., CLARK, R. & DEHAIS, P. 2009. Design of a high density SNP genotyping assay in the pig using SNPs Identified and characterized by next generation sequencing technology. *PLoS One,* 4.
RAMSEY, L. B., JANKE, L. J., EDICK, M. J., CHENG, C., WILLIAMS, R. T., SHERR, C. J., EVANS, W. E. & RELLING, M. V. 2014. Host Thiopurine Methyltransferase Status Affects Mercaptopurine Antileukemic Effectiveness in a Murine Model. *Pharmacogenetics and genomics,* 24**,** 263-271.
RATHJE, T. A., ROHRER, G. A. & JOHNSON, R. K. 1997. Evidence for quantitative trait loci affecting ovulation rate in pigs. *J Anim Sci,* 75**,** 1486-94.
RAU, V., IYER, S. V., OH, I., CHANDRA, D., HARRISON, N., EGER, E. I., FANSELOW, M. S., HOMANICS, G. E. & SONNER, J. M. 2009. GABA(A) Receptor Alpha 4 Subunit Knockout Mice Are Resistant to the Amnestic Effect of Isoflurane. *Anesthesia and analgesia,* 109**,** 1816-1822.
REBER, I., KELLER, I., BECKER, D., FLURY, C., WELLE, M. & DRÖGEMÜLLER, C. 2015. Wattles in goats are associated with the FMN1/GREM1 region on chromosome 10. *Animal Genetics,* 46**,** 316-320.
RENTHAL, N. E., CHEN, C.-C., WILLIAMS, K. R. C., GERARD, R. D., PRANGE-KIEL, J. & MENDELSON, C. R. 2010. miR-200 family and targets, ZEB1 and ZEB2, modulate uterine quiescence and contractility during pregnancy and labor. *Proceedings of the National Academy of Sciences,* 107**,** 20828.
RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13**,** 278-289.
RHODES, J., BEALE, M. A. & FISHER, M. C. 2014. Illuminating Choices for Library Prep: A Comparison of Library Preparation Methods for Whole Genome Sequencing of Cryptococcus neoformans Using Illumina HiSeq. *PLOS ONE,* 9**,** e113501.
RIMMER, A., PHAN, H., MATHIESON, I., IQBAL, Z., TWIGG, S. R., CONSORTIUM, W. G. S., WILKIE, A. O., MCVEAN, G. & LUNTER, G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet,* 46**,** 912-8.
RINGWALD, M., EPPIG, J. T., KADIN, J. A., RICHARDSON, J. E. & THE GENE EXPRESSION DATABASE, G. 2000. GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Research,* 28**,** 115-119.

RISSE, J., THOMSON, M., PATRICK, S., BLAKELY, G., KOUTSOVOULOS, G., BLAXTER, M. & WATSON, M. 2015. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *GigaScience,* 4**,** 60.
RITCHIE, G. R. & FLICEK, P. 2014. Computational approaches to interpreting genomic sequence variation. *Genome Med,* 6**,** 87.
RITCHIE, G. R. S., DUNHAM, I., ZEGGINI, E. & FLICEK, P. 2014. Functional annotation of noncoding sequence variants. *Nat Meth,* 11**,** 294-296.
ROBERT, C., FUENTES-UTRILLA, P., TROUP, K., LOECHERBACH, J., TURNER, F., TALBOT, R., ARCHIBALD, A. L., MILEHAM, A., DEEB, N., HUME, D. A. & WATSON, M. 2014. Design and development of exome capture sequencing for the domestic pig (Sus scrofa). *Bmc Genomics,* 15**,** 9.
ROBERTS, R. J., CARNEIRO, M. O. & SCHATZ, M. C. 2013. The advantages of SMRT sequencing. *Genome Biology,* 14**,** 405-405.
ROBINSON, J. T., THORVALDSDOTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nat Biotech,* 29**,** 24-26.
ROBINSON, P. N. 2012. Deep phenotyping for precision medicine. *Hum Mutat,* 33**,** 777-80.
RODMAN, D. M. & ZAMUDIO, S. 1991. The cystic fibrosis heterozygote--advantage in surviving cholera? *Med Hypotheses,* 36**,** 253-8.
ROGG, M., YASUDA-YAMAHARA, M., ABED, A., DINSE, P., HELMSTÄDTER, M., CONZELMANN, A. C., FRIMMEL, J., SELLUNG, D., BINIOSSEK, M. L., KRETZ, O., GRAHAMMER, F., SCHILLING, O., HUBER, T. B. & SCHELL, C. 2017. The WD40-domain containing protein CORO2B is specifically enriched in glomerular podocytes and regulates the ventral actin cytoskeleton. *Scientific Reports,* 7**,** 15910.
ROGOZIN, I. B. & MILANESI, L. 1997. Analysis of donor splice sites in different eukaryotic organisms. *Journal of Molecular Evolution,* 45**,** 50-59.
ROHLAND, N. & REICH, D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res,* 22**,** 939-46.
ROHRER, G. A., ALEXANDER, L. J., HU, Z., SMITH, T. P., KEELE, J. W. & BEATTIE, C. W. 1996. A comprehensive map of the porcine genome. *Genome Res,* 6**,** 371-91.
ROMANOVA, L. G., ANGER, M., ZATSEPINA, O. V. & SCHULTZ, R. M. 2006. Implication of Nucleolar Protein SURF6 in Ribosome Biogenesis and Preimplantation Mouse Development1. *Biology of Reproduction,* 75**,** 690-696.
ROS-FREIXEDES, R., GONEN, S., GORJANC, G. & HICKEY, J. M. 2017. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. *Genetics, selection, evolution : GSE* [Online], 49. Available: https://doi.org/10.1186/s12711-017-0353-y [Accessed 2017/10//].

ROSSI, J., HERZIG, K.-H., VÕIKAR, V., HILTUNEN, P. H., SEGERSTRÅLE, M. & AIRAKSINEN, M. S. 2003. Alimentary tract innervation deficits and dysfunction in mice lacking GDNF family receptor α2. *The Journal of Clinical Investigation,* 112**,** 707-716.

ROTHSCHILD, M., JACOBSON, C., VASKE, D., TUGGLE, C., WANG, L., SHORT, T., ECKARDT, G., SASAKI, S., VINCENT, A., MCLAREN, D., SOUTHWOOD, O., VAN DER STEEN, H., MILEHAM, A. & PLASTOW, G. 1996. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Sciences,* 93**,** 201.

RUAN, J., XU, J., CHEN-TSAI, R. Y. & LI, K. 2017. Genome editing in livestock: Are we ready for a revolution in animal breeding industry? *Transgenic Research,* 26**,** 715-726.

RULL, K., NAGIRNAJA, L. & LAAN, M. 2012. Genetics of Recurrent Miscarriage: Challenges, Current Knowledge, Future Directions. *Frontiers in Genetics,* 3**,** 34.

SADEDIN, S., ELLIS, J., MASTERS, S. & OSHLACK, A. 2018. Ximmer: A System for Improving Accuracy and Consistency of CNV Calling from Exome Data. *bioRxiv*.

SAHLIN, K., VEZZI, F., NYSTEDT, B., LUNDEBERG, J. & ARVESTAD, L. 2014. BESST--efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics,* 15.

SAIKI, R. K., GELFAND, D. H., STOFFEL, S., SCHARF, S. J., HIGUCHI, R., HORN, G. T., MULLIS, K. B. & ERLICH, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science,* 239**,** 487.

SALMELA, L., MAKINEN, V., VALIMAKI, N., YLINEN, J. & UKKONEN, E. 2011. Fast scaffolding with small independent mixed integer programs. *Bioinformatics,* 27.

SALMELA, L. & RIVALS, E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics,* 30**,** 3506-3514.

SALZBERG, S. L. & YORKE, J. A. 2005. Beware of mis-assembled genomes. *Bioinformatics,* 21**,** 4320-4321.

SAN MARTIN, S., SOTO-SUAZO, M. & ZORN, T. M. T. 2003. Distribution of versican and hyaluronan in the mouse uterus during decidualization. *Brazilian Journal of Medical and Biological Research,* 36**,** 1067-1071.

SANGER, F. & COULSON, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94**,** 441-8.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-5467.

SASSI, C., GUERREIRO, R., GIBBS, R., DING, J., LUPTON, M. K., TROAKES, C., LUNNON, K., AL-SARRAJ, S., BROWN, K. S., MEDWAY, C., LORD, J., TURTON, J., MANN, D., SNOWDEN, J., NEARY, D., HARRIS, J., BRAS, J., CONSORTIUM, A., MORGAN, K., POWELL, J. F., SINGLETON, A. & HARDY, J. 2014. Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and

p.S230R) in British early-onset Alzheimer's disease. *Neurobiol Aging,* 35**,** 2422 e13-6.
SATHIRAPONGSASUTI, J. F., LEE, H., HORST, B. A. J., BRUNNER, G., COCHRAN, A. J., BINDER, S., QUACKENBUSH, J. & NELSON, S. F. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics,* 27**,** 2648-2654.
SAYYAB, S., VILUMA, A., BERGVALL, K., BRUNBERG, E., JAGANNATHAN, V., LEEB, T., ANDERSSON, G. & BERGSTRÖM, T. F. 2016. Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis. *G3: Genes|Genomes|Genetics,* 6**,** 521-527.
SBONER, A., MU, X. J., GREENBAUM, D., AUERBACH, R. K. & GERSTEIN, M. B. 2011. The real cost of sequencing: higher than you think! *Genome Biology,* 12**,** 125-125.
SCHATZ, M. C. 2018. In Pursuit of Perfect Genome Sequencing: Accurate Identification and Phasing of Structural Variations using Short, Long, and Linked Reads. *PAG XXVI.* San Diego.
SCHATZ, M. C., DELCHER, A. L. & SALZBERG, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research,* 20**,** 1165-1173.
SCHAUB, M. A., BOYLE, A. P., KUNDAJE, A., BATZOGLOU, S. & SNYDER, M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res,* 22**,** 1748-59.
SCHICK, U. M., AUER, P. L., BIS, J. C., LIN, H., WEI, P., PANKRATZ, N., LANGE, L. A., BRODY, J., STITZIEL, N. O., KIM, D. S., CARLSON, C. S., FORNAGE, M., HAESSLER, J., HSU, L., JACKSON, R. D., KOOPERBERG, C., LEAL, S. M., PSATY, B. M., BOERWINKLE, E., TRACY, R., ARDISSINO, D., SHAH, S., WILLER, C., LOOS, R., MELANDER, O., MCPHERSON, R., HOVINGH, K., REILLY, M., WATKINS, H., GIRELLI, D., FONTANILLAS, P., CHASMAN, D. I., GABRIEL, S. B., GIBBS, R., NICKERSON, D. A., KATHIRESAN, S., PETERS, U., DUPUIS, J., WILSON, J. G., RICH, S. S., MORRISON, A. C., BENJAMIN, E. J., GROSS, M. D., REINER, A. P., COHORTS FOR, H., AGING RESEARCH IN GENOMIC, E., THE NATIONAL HEART, L. & BLOOD INSTITUTE, G. O. E. S. P. 2015. Association of exome sequences with plasma C-reactive protein levels in >9000 participants. *Hum Mol Genet,* 24**,** 559-71.
SCHNEIDER, J. F., MILES, J. R., BROWN-BRANDL, T. M., NIENABER, J. A., ROHRER, G. A. & VALLET, J. L. 2015. Genomewide association analysis for average birth interval and stillbirth in swine. *J Anim Sci,* 93**,** 529-40.
SCHNEIDER, J. F., REMPEL, L. A. & ROHRER, G. A. 2012. Genome-wide association study of swine farrowing traits. Part I: Genetic and genomic parameter estimates1,2. *Journal of Animal Science,* 90**,** 3353-3359.
SCHOOK, L. B., BEEVER, J. E., ROGERS, J., HUMPHRAY, S., ARCHIBALD, A. & CHARDON, P. 2005. Swine Genome Sequencing

Consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comp Funct Genomics,* 6.
SCHUBERT, M., GINOLHAC, A., LINDGREEN, S., THOMPSON, J. F., AL-RASHEID, K. A., WILLERSLEV, E., KROGH, A. & ORLANDO, L. 2012. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics,* 13**,** 178.
SCHWARZ, J. M., RODELSPERGER, C., SCHUELKE, M. & SEELOW, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth,* 7**,** 575-576.
SEDLAZECK, F. J., RESCHENEDER, P., SMOLKA, M., FANG, H., NATTESTAD, M., VON HAESELER, A. & SCHATZ, M. C. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*.
SEO, H., KIM, M., CHOI, Y., CHOI, Y. & KA, H. 2011. Expression and Regulation of IL1B Receptor Subtypes, IL1R1 and IL1RAP, in the Uterine Endometrium During Pregnancy in Pigs. *Biology of Reproduction,* 85**,** 484-484.
SERVIN, B., FARAUT, T., IANNUCCELLI, N., ZELENIKA, D. & MILAN, D. 2012. High-resolution autosomal radiation hybrid maps of the pig genome and their contribution to the genome sequence assembly. *BMC Genomics,* 13**,** 585.
SHAHEEN, R., SCHMIDTS, M., FAQEIH, E., HASHEM, A., LAUSCH, E., HOLDER, I., SUPERTI-FURGA, A., MITCHISON, H. M., ALMOISHEER, A., ALAMRO, R., ALSHIDDI, T., ALZAHRANI, F., BEALES, P. L. & ALKURAYA, F. S. 2015. A founder CEP120 mutation in Jeune asphyxiating thoracic dystrophy expands the role of centriolar proteins in skeletal ciliopathies. *Human Molecular Genetics,* 24**,** 1410-1419.
SHAN, N., ZHANG, X., XIAO, X., ZHANG, H., TONG, C., LUO, X., CHEN, Y., LIU, X., YIN, N., DENG, Q. & QI, H. 2015. Laminin (LAMA4) expression promotes trophoblast cell invasion, migration, and angiogenesis, and is lowered in preeclamptic placentas. *Placenta,* 36**,** 809-820.
SHARON, D., TILGNER, H., GRUBERT, F. & SNYDER, M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotech,* 31**,** 1009-1014.
SHEPHERD, R. K. & KINGHORN, B. P. 1992. Optimising multi-tier open nucleus breeding schemes. *Theoretical and Applied Genetics,* 85**,** 372-378.
SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research,* 29**,** 308-311.
SHI, Y., LI, Y., ZHANG, D., ZHANG, H., LI, Y., LU, F., LIU, X., HE, F., GONG, B., CAI, L., LI, R., LIAO, S., MA, S., LIN, H., CHENG, J., ZHENG, H., SHAN, Y., CHEN, B., HU, J., JIN, X., ZHAO, P., CHEN, Y., ZHANG, Y., LIN, Y., LI, X., FAN, Y., YANG, H., WANG, J. & YANG, Z. 2011. Exome Sequencing Identifies ZNF644 Mutations in High Myopia. *PLoS Genet,* 7**,** e1002084.

SHIN, S. C., AHN, D. H., KIM, S. J., LEE, H., OH, T.-J., LEE, J. E. & PARK, H. 2013. Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS ONE,* 8**,** e68824.
SHORT, T. H., ROTHSCHILD, M. F., SOUTHWOOD, O. I., MCLAREN, D. G., DE VRIES, A., VAN DER STEEN, H., ECKARDT, G. R., TUGGLE, C. K., HELM, J., VASKE, D. A., MILEHAM, A. J. & PLASTOW, G. S. 1997. Effect of the estrogen receptor locus on reproduction and production traits in four commercial pig lines. *J Anim Sci,* 75**,** 3138-42.
SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M., ROSENBLOOM, K., CLAWSON, H., SPIETH, J., HILLIER, L. W., RICHARDS, S., WEINSTOCK, G. M., WILSON, R. K., GIBBS, R. A., KENT, W. J., MILLER, W. & HAUSSLER, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research,* 15**,** 1034-1050.
SILVA, G. G. Z., DUTILH, B. E., MATTHEWS, T. D., ELKINS, K., SCHMIEDER, R., DINSDALE, E. A. & EDWARDS, R. A. 2013. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code for Biology and Medicine,* 8**,** 23-23.
SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31**,** 3210-3212.
SIMPSON, J. T. & DURBIN, R. 2011. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*.
SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J. & BIROL, I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res,* 19.
SINDI, S., HELMAN, E., BASHIR, A. & RAPHAEL, B. J. 2009. A geometric approach for classification and comparison of structural variants. *Bioinformatics,* 25**,** i222-i230.
SKINNER, B. M., SARGENT, C. A., CHURCHER, C., HUNT, T., HERRERO, J. & LOVELAND, J. E. 2016. The pig X and Y Chromosomes: structure, sequence, and evolution. *Genome Res,* 26.
SMEDLEY, D., HAIDER, S., DURINCK, S., PANDINI, L., PROVERO, P., ALLEN, J., ARNAIZ, O., AWEDH, M. H., BALDOCK, R., BARBIERA, G., BARDOU, P., BECK, T., BLAKE, A., BONIERBALE, M., BROOKES, A. J., BUCCI, G., BUETTI, I., BURGE, S., CABAU, C., CARLSON, J. W., CHELALA, C., CHRYSOSTOMOU, C., CITTARO, D., COLLIN, O., CORDOVA, R., CUTTS, R. J., DASSI, E., GENOVA, A. D., DJARI, A., ESPOSITO, A., ESTRELLA, H., EYRAS, E., FERNANDEZ-BANET, J., FORBES, S., FREE, R. C., FUJISAWA, T., GADALETA, E., GARCIA-MANTEIGA, J. M., GOODSTEIN, D., GRAY, K., GUERRA-ASSUNÇÃO, J. A., HAGGARTY, B., HAN, D.-J., HAN, B. W., HARRIS, T., HARSHBARGER, J., HASTINGS, R. K., HAYES, R. D., HOEDE, C., HU, S., HU, Z.-L., HUTCHINS, L., KAN, Z., KAWAJI, H., KELIET, A., KERHORNOU, A., KIM, S., KINSELLA, R., KLOPP, C., KONG, L., LAWSON, D., LAZAREVIC, D., LEE, J.-H., LETELLIER, T., LI, C.-Y., LIO, P., LIU, C.-J., LUO, J., MAASS, A.,

MARIETTE, J., MAUREL, T., MERELLA, S., MOHAMED, A. M., MOREEWS, F., NABIHOUDINE, I., NDEGWA, N., NOIROT, C., PEREZ-LLAMAS, C., PRIMIG, M., QUATTRONE, A., QUESNEVILLE, H., RAMBALDI, D., REECY, J., RIBA, M., ROSANOFF, S., SADDIQ, A. A., SALAS, E., SALLOU, O., SHEPHERD, R., SIMON, R., SPERLING, L., SPOONER, W., STAINES, D. M., STEINBACH, D., STONE, K., STUPKA, E., TEAGUE, J. W., DAYEM ULLAH, A. Z., WANG, J., WARE, D., et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research,* 43**,** W589-W598.

SMIT, A., HUBLEY, R & GREEN, P. . 2013-2015. *RepeatMasker Open-4.0.* [Online]. Available: http://www.repeatmasker.org [Accessed 16/05/2016].

SMITAL, J., WOLF, J. & DE SOUSA, L. L. 2005. Estimation of genetic parameters of semen characteristics and reproductive traits in AI boars. *Anim Reprod Sci,* 86**,** 119-30.

SMITH, H. E. & YUN, S. 2017. Evaluating alignment and variant-calling software for mutation identification in C. elegans by whole-genome sequencing. *PLoS ONE,* 12**,** e0174446.

SMITH, J. D., HING, A. V., CLARKE, C. M., JOHNSON, N. M., PEREZ, F. A., PARK, S. S., HORST, J. A., MECHAM, B., MAVES, L., NICKERSON, D. A., UNIVERSITY OF WASHINGTON CENTER FOR MENDELIAN, G. & CUNNINGHAM, M. L. 2014. Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontonasal dysostosis. *Am J Hum Genet,* 95**,** 235-40.

SNAPE, K., RUARK, E., TARPEY, P., RENWICK, A., TURNBULL, C., SEAL, S., MURRAY, A., HANKS, S., DOUGLAS, J., STRATTON, M. R. & RAHMAN, N. 2012. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res Treat,* 134**,** 429-33.

SNYDER, M., DU, J. & GERSTEIN, M. 2010. Personal genome sequencing: current approaches and challenges. *Genes & Development,* 24**,** 423-431.

SOUTHERN, E. M. 1982. New methods for analysing DNA make genetics simpler. *Biochemical Society Transactions,* 10**,** 1.

SOVIĆ, I., ŠIKIĆ, M., WILM, A., FENLON, S. N., CHEN, S. & NAGARAJAN, N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications,* 7**,** 11307.

SPENCER, C. C. A., SU, Z., DONNELLY, P. & MARCHINI, J. 2009. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLOS Genetics,* 5**,** e1000477.

STADEN, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res,* 6**,** 2601-10.

STANKIEWICZ, P. & LUPSKI, J. R. 2010. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine,* 61**,** 437-455.

STAR, B. & SPENCER, H. G. 2013. Effects of Genetic Drift and Gene Flow on the Selective Maintenance of Genetic Variation. *Genetics,* 194**,** 235-244.
STEFL, S., NISHI, H., PETUKH, M., PANCHENKO, A. R. & ALEXOV, E. 2013. Molecular mechanisms of disease-causing missense mutations. *Journal of molecular biology,* 425**,** 3919-3936.
STENHOUSE, C. 2017. Integrin subunit expression in porcine endometrial tissue supplying small and normal-sized foetuses throughout gestation. *Ferility.* Edinburgh, UK.
STRANGER, B. E., STAHL, E. A. & RAJ, T. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics,* 187**,** 367-383.
SU, Z.-J., HAHN, C. N., GOODALL, G. J., RECK, N. M., LESKE, A. F., DAVY, A., KREMMIDIOTIS, G., VADAS, M. A. & GAMBLE, J. R. 2004. A vascular cell-restricted RhoGAP, p73RhoGAP, is a key regulator of angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 12212.
SULSTON, J., DU, Z., THOMAS, K., WILSON, R., HILLIER, L., STADEN, R., HALLORAN, N., GREEN, P., THIERRY-MIEG, J., QIU, L. & ET AL. 1992. The C. elegans genome sequencing project: a beginning. *Nature,* 356**,** 37-41.
SUN, H. F., ERNST, C. W., YERLE, M., PINTON, P., ROTHSCHILD, M. F., CHARDON, P., ROGEL-GAILLARD, C. & TUGGLE, C. K. 1999. Human chromosome 3 and pig chromosome 13 show complete synteny conservation but extensive gene-order differences. *Cytogenet Cell Genet,* 85**,** 273-8.
SUN, Z., CERABONA, D., HE, Y. & NALEPA, G. 2016. Cdkn3 Knockout Mice Develop Hematopoietic Malignancies. *Blood,* 128**,** 1537-1537.
SUZUKI, S., YASUDA, T., SHIRAISHI, Y., MIYANO, S. & NAGASAKI, M. 2011. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics,* 12**,** S7.
SWINDLE, M. M., MAKIN, A., HERRON, A. J., CLUBB, F. J. & FRAZIER, K. S. 2011. Swine as Models in Biomedical Research and Toxicology Testing. *Veterinary Pathology,* 49**,** 344-356.
TABANGIN, M. E., WOO, J. G. & MARTIN, L. J. 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings,* 3**,** S41-S41.
TAKENOUCHI, T., TSUKAHARA, Y., HORIKAWA, R., KOSAKI, K. & KOSAKI, R. 2014. Four-decade-old mummified umbilical tissue making retrospective molecular diagnosis of ornithine carbamoyltransferase deficiency. *American Journal of Medical Genetics Part A,* 164**,** 2679-2681.
TALLA, V., SUH, A., KALSOOM, F., DINCĂ, V., VILA, R., FRIBERG, M., WIKLUND, C. & BACKSTRÖM, N. 2017. Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in

Wood-White (Leptidea) Butterflies. *Genome Biology and Evolution,* 9**,** 2491-2505.
TANERI, B., ASILMAZ, E. & GAASTERLAND, T. 2012. Biomedical impact of splicing mutations revealed through exome sequencing. *Mol Med,* 18**,** 314-9.
TANN, A. W., BOLDOGH, I., MEISS, G., QIAN, W., VAN HOUTEN, B., MITRA, S. & SZCZESNY, B. 2011. Apoptosis Induced by Persistent Single-strand Breaks in Mitochondrial Genome: CRITICAL ROLE OF EXOG (5′-EXO/ENDONUCLEASE) IN THEIR REPAIR. *The Journal of Biological Chemistry,* 286**,** 31975-31983.
TENA-BETANCOURT, E., TENA-BETANCOURT, C. A., ZÚNIGA-MUÑOZ, A. M., HERNÁNDEZ-GODÍNEZ, B., IBÁÑEZ-CONTRERAS, A. & GRAULLERA-RIVERA, V. 2014. Multiple Extrauterine Pregnancy with Early and Near Full-Term Mummified Fetuses in a New Zealand White Rabbit (Oryctolagus cuniculus). *Journal of the American Association for Laboratory Animal Science : JAALAS,* 53**,** 204-207.
TER BRAAK, C. J. F., BOER, M. P. & BINK, M. C. A. M. 2005. Extending Xu's Bayesian Model for Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics,* 170**,** 1435.
TERPSTRA, C., WENSVOORT, G. & POL, J. M. 1991. Experimental reproduction of porcine epidemic abortion and respiratory syndrome (mystery swine disease) by infection with Lelystad virus: Koch's postulates fulfilled. *Vet Q,* 13**,** 131-6.
THAN, B. L., LINNEKAMP, J. F., STARR, T. K., LARGAESPADA, D. A., ROD, A., ZHANG, Y., BRUNER, V., ABRAHANTE, J., SCHUMANN, A., LUCZAK, T., NIEMCZYK, A., O'SULLIVAN, M. G., MEDEMA, J. P., FIJNEMAN, R. J., MEIJER, G. A., VAN DEN BROEK, E., HODGES, C. A., SCOTT, P. M., VERMEULEN, L. & CORMIER, R. T. 2016. CFTR is a tumor suppressor gene in murine and human intestinal cancer. *Oncogene,* 35**,** 4179-87.
THE FANTOM CONSORTIUM, THE RIKEN PMI & CLST 2014. A promoter-level mammalian expression atlas. *Nature,* 507**,** 462-470.
THE INTERNATIONAL BARLEY GENOME SEQUENCING CONSORTIUM 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature,* 491**,** 711-716.
THORSEN, K., SORENSEN, K. D., BREMS-ESKILDSEN, A. S., MODIN, C., GAUSTADNES, M., HEIN, A. M., KRUHOFFER, M., LAURBERG, S., BORRE, M., WANG, K., BRUNAK, S., KRAINER, A. R., TORRING, N., DYRSKJOT, L., ANDERSEN, C. L. & ORNTOFT, T. F. 2008. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol Cell Proteomics,* 7**,** 1214-24.
THORVALDSDÓTTIR, H., ROBINSON, J. T. & MESIROV, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics,* 14**,** 178-192.
TIGCHELAAR, W., DE JONG, A. M., VAN GILST, W. H., DE BOER, R. A. & SILLJÉ, H. H. W. 2016. In EXOG-depleted cardiomyocytes cell death is marked by a decreased mitochondrial reserve capacity of the electron transport chain. *BioEssays,* 38**,** S136-S145.

TIGCHELAAR, W., YU, H., DE JONG, A. M., VAN GILST, W. H., VAN DER HARST, P., WESTENBRINK, B. D., DE BOER, R. A. & SILLJÉ, H. H. W. 2014. Loss of mitochondrial exo/endonuclease EXOG affects mitochondrial respiration and induces ROS-mediated cardiomyocyte hypertrophy. *American Journal of Physiology-Cell Physiology,* 308**,** C155-C163.

TORTEREAU, F., SERVIN, B., FRANTZ, L., MEGENS, H. J., MILAN, D., ROHRER, G., WIEDMANN, R., BEEVER, J., ARCHIBALD, A. L., SCHOOK, L. B. & GROENEN, M. A. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics, accepted for publ.*

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

TRUBETSKOY, V., RODRIGUEZ, A., DAVE, U., CAMPBELL, N., CRAWFORD, E. L., COOK, E. H., SUTCLIFFE, J. S., FOSTER, I., MADDURI, R., COX, N. J. & DAVIS, L. K. 2015. Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes. *Bioinformatics,* 31**,** 187-93.

TSUCHIYA, Y., YOSHIBA, S., GUPTA, A., WATANABE, K. & KITAGAWA, D. 2016. Cep295 is a conserved scaffold protein required for generation of a bona fide mother centriole. *Nature Communications,* 7**,** 12567.

TUSKAN, G. A., DIFAZIO, S., JANSSON, S., BOHLMANN, J., GRIGORIEV, I., HELLSTEN, U., PUTNAM, N., RALPH, S., ROMBAUTS, S., SALAMOV, A., SCHEIN, J., STERCK, L., AERTS, A., BHALERAO, R. R., BHALERAO, R. P., BLAUDEZ, D., BOERJAN, W., BRUN, A., BRUNNER, A., BUSOV, V., CAMPBELL, M., CARLSON, J., CHALOT, M., CHAPMAN, J., CHEN, G.-L., COOPER, D., COUTINHO, P. M., COUTURIER, J., COVERT, S., CRONK, Q., CUNNINGHAM, R., DAVIS, J., DEGROEVE, S., DÉJARDIN, A., DEPAMPHILIS, C., DETTER, J., DIRKS, B., DUBCHAK, I., DUPLESSIS, S., EHLTING, J., ELLIS, B., GENDLER, K., GOODSTEIN, D., GRIBSKOV, M., GRIMWOOD, J., GROOVER, A., GUNTER, L., HAMBERGER, B., HEINZE, B., HELARIUTTA, Y., HENRISSAT, B., HOLLIGAN, D., HOLT, R., HUANG, W., ISLAM-FARIDI, N., JONES, S., JONES-RHOADES, M., JORGENSEN, R., JOSHI, C., KANGASJÄRVI, J., KARLSSON, J., KELLEHER, C., KIRKPATRICK, R., KIRST, M., KOHLER, A., KALLURI, U., LARIMER, F., LEEBENS-MACK, J., LEPLÉ, J.-C., LOCASCIO, P., LOU, Y., LUCAS, S., MARTIN, F., MONTANINI, B., NAPOLI, C., NELSON, D. R., NELSON, C., NIEMINEN, K., NILSSON, O., PEREDA, V., PETER, G., PHILIPPE, R., PILATE, G., POLIAKOV, A., RAZUMOVSKAYA, J., RICHARDSON, P., RINALDI, C., RITLAND, K., ROUZÉ, P., RYABOY, D., SCHMUTZ, J., SCHRADER, J., SEGERMAN, B., SHIN, H., SIDDIQUI, A., STERKY, F., TERRY, A., TSAI, C.-J., UBERBACHER,

E., UNNEBERG, P., et al. 2006. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science,* 313**,** 1596-1604.
VALLENDER, E. J. 2011. Expanding whole exome resequencing into non-human primates. *Genome Biol,* 12**,** R87.
VAN DEN BERG, I., BOICHARD, D., GULDBRANDTSEN, B. & LUND, M. S. 2016. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. *G3: Genes|Genomes|Genetics,* 6**,** 2553-2561.
VAN DER LENDE, T. & VAN RENS, B. T. T. M. 2003. Critical periods for foetal mortality in gilts identified by analysing the length distribution of mummified foetuses and frequency of non-fresh stillborn piglets. *Animal Reproduction Science,* 75**,** 141-150.
VAN SON, M., ENGER, E. G., GROVE, H., ROS-FREIXEDES, R., KENT, M. P., LIEN, S. & GRINDFLEK, E. 2017. Genome-wide association study confirm major QTL for backfat fatty acid composition on SSC14 in Duroc pigs. *BMC Genomics,* 18**,** 369.
VARSHNEY, M. K., INZUNZA, J., LUPU, D., GANAPATHY, V., ANTONSON, P., RÜEGG, J., NALVARTE, I. & GUSTAFSSON, J.-Å. 2017. Role of estrogen receptor beta in neural differentiation of mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*.
VASER, R., SOVIC, I., NAGARAJAN, N. & SIKIC, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res,* 27**,** 737-746.
VERHOEVEN, K., DE JONGHE, P., VAN DE PUTTE, T., NELIS, E., ZWIJSEN, A., VERPOORTEN, N., DE VRIENDT, E., JACOBS, A., VAN GERWEN, V., FRANCIS, A., CEUTERICK, C., HUYLEBROECK, D. & TIMMERMAN, V. 2003. Slowed Conduction and Thin Myelination of Peripheral Nerves Associated with Mutant Rho Guanine-Nucleotide Exchange Factor 10. *The American Journal of Human Genetics,* 73**,** 926-932.
VERONEZE, R., LOPES, P. S., GUIMARAES, S. E., SILVA, F. F., LOPES, M. S., HARLIZIUS, B. & KNOL, E. F. 2013. Linkage disequilibrium and haplotype block structure in six commercial pig lines. *J Anim Sci,* 91**,** 3493-501.
VERTEBRATE GENOMES PROJECT. 2018. *Vertebrate Genomes Project.*
VIGNEAULT, C., GRAVEL, C., VALLÉE, M., MCGRAW, S. & SIRARD, M.-A. 2009. Unveiling the bovine embryo transcriptome during the maternal-to-embryonic transition. *Reproduction,* 137**,** 245-257.
VILLUMSEN, T. M., JANSS, L. & LUND, M. S. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics,* 126**,** 3-13.
VISSCHER, P., PONG-WONG, R., WHITTEMORE, C. & HALEY, C. 2000. Impact of biotechnology on (cross)breeding programmes in pigs. *Livestock Production Science,* 65**,** 57-70.
VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. & YANG, J. 2017. 10 Years of GWAS Discovery:

Biology, Function, and Translation. *American Journal of Human Genetics,* 101**,** 5-22.
VOYTAS, D. F. & GAO, C. 2014. Precision Genome Engineering and Agriculture: Opportunities and Regulatory Challenges. *PLOS Biology,* 12**,** e1001877.
VRBACKÝ, M., KOVALČÍKOVÁ, J., CHAWENGSAKSOPHAK, K., BECK, I. M., MRÁČEK, T., NŮSKOVÁ, H., SEDMERA, D., PAPOUŠEK, F., KOLÁŘ, F., SOBOL, M., HOZÁK, P., SEDLACEK, R. & HOUŠTĚK, J. 2016. Knockout of Tmem70 alters biogenesis of ATP synthase and leads to embryonal lethality in mice. *Human Molecular Genetics,* 25**,** 4674-4685.
WADE, C. M., GIULOTTO, E., SIGURDSSON, S., ZOLI, M., GNERRE, S., IMSLAND, F., LEAR, T. L., ADELSON, D. L., BAILEY, E., BELLONE, R. R., BLOCKER, H., DISTL, O., EDGAR, R. C., GARBER, M., LEEB, T., MAUCELI, E., MACLEOD, J. N., PENEDO, M. C., RAISON, J. M., SHARPE, T., VOGEL, J., ANDERSSON, L., ANTCZAK, D. F., BIAGI, T., BINNS, M. M., CHOWDHARY, B. P., COLEMAN, S. J., DELLA VALLE, G., FRYC, S., GUERIN, G. & ET, A. L. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science,* 326.
WALKER, B. J., ABEEL, T., SHEA, T., PRIEST, M., ABOUELLIEL, A., SAKTHIKUMAR, S., CUOMO, C. A., ZENG, Q., WORTMAN, J., YOUNG, S. K. & EARL, A. M. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE,* 9**,** e112963.
WALSH, J. B. & MARKS, J. 1986. Sequencing the human genome. *Nature,* 322**,** 590.
WANG, H., CHEN, X., DUDINSKY, L., PATENIA, C., CHEN, Y., LI, Y., WEI, Y., ABBOUD, E. B., AL-RAJHI, A. A., LEWIS, R. A., LUPSKI, J. R., MARDON, G., GIBBS, R. A., PERKINS, B. D. & CHEN, R. 2011a. Exome capture sequencing identifies a novel mutation in BBS4. *Molecular Vision,* 17**,** 3529-3540.
WANG, J., MULLIGHAN, C. G., EASTON, J., ROBERTS, S., HEATLEY, S. L., MA, J., RUSCH, M. C., CHEN, K., HARRIS, C. C., DING, L., HOLMFELDT, L., PAYNE-TURNER, D., FAN, X., WEI, L., ZHAO, D., OBENAUER, J. C., NAEVE, C., MARDIS, E. R., WILSON, R. K., DOWNING, J. R. & ZHANG, J. 2011b. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Meth,* 8**,** 652-654.
WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.
WAPLES, R. S. 2015. Testing for Hardy–Weinberg Proportions: Have We Lost the Plot? *Journal of Heredity,* 106**,** 1-19.
WARD, L. D. & KELLIS, M. 2012. Interpreting non-coding variation in complex disease genetics. *Nature biotechnology,* 30**,** 1095-1106.

WARDEN, C. D., ADAMSON, A. W., NEUHAUSEN, S. L. & WU, X. 2014. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ,* 2**,** e600.
WARR, A., ROBERT, C., HUME, D., ARCHIBALD, A., DEEB, N. & WATSON, M. 2015a. Exome Sequencing: Current and Future Perspectives. *G3: Genes|Genomes|Genetics*.
WARR, A., ROBERT, C., HUME, D., ARCHIBALD, A. L., DEEB, N. & WATSON, M. 2015b. Identification of low-confidence regions in the pig reference genome (Sscrofa10.2). *Frontiers in Genetics,* 6.
WATSON, M. 2014. Illuminating the future of DNA sequencing. *Genome Biology,* 15**,** 108-108.
WATSON, M. 2018. Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. *bioRxiv*.
WATSON, M. & WARR, A. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology,* 37**,** 124-126.
WEI, Z., WANG, W., HU, P., LYON, G. J. & HAKONARSON, H. 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research,* 39**,** e132-e132.
WEIGEL, K. A., DE LOS CAMPOS, G., GONZÁLEZ-RECIO, O., NAYA, H., WU, X. L., LONG, N., ROSA, G. J. M. & GIANOLA, D. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science,* 92**,** 5248-5257.
WEISENFELD, N. I., YIN, S., SHARPE, T., LAU, B., HEGARTY, R., HOLMES, L., SOGOLOFF, B., TABBAA, D., WILLIAMS, L., RUSS, C., NUSBAUM, C., LANDER, E. S., MACCALLUM, I. & JAFFE, D. B. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet,* 46**,** 1350-5.
WELLS, K. D. & PRATHER, R. S. 2017. Genome-editing technologies to improve research, reproduction, and production in pigs. *Molecular Reproduction and Development,* 84**,** 1012-1017.
WELSH, M. J., ROGERS, C. S., STOLTZ, D. A., MEYERHOLZ, D. K. & PRATHER, R. S. 2009. Development of a Porcine Model of Cystic Fibrosis. *Transactions of the American Clinical and Climatological Association,* 120**,** 149-162.
WENDLER, N., MASCHER, M., NOH, C., HIMMELBACH, A., SCHOLZ, U., RUGE-WEHLING, B. & STEIN, N. 2014. Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol J,* 12**,** 1122-31.
WERNERSSON, R., SCHIERUP, M. H., JORGENSEN, F. G., GORODKIN, J., PANITZ, F., STAERFELDT, H. H., CHRISTENSEN, O. F., MAILUND, T., HORNSHOJ, H., KLEIN, A., WANG, J., LIU, B., HU, S., DONG, W., LI, W., WONG, G. K., YU, J., WANG, J., BENDIXEN, C., FREDHOLM, M., BRUNAK, S., YANG, H. & BOLUND, L. 2005. Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing. *BMC Genomics,* 6**,** 70.

WETTERSTRAND, K. A. 2017. *Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* [Online]. Available: www.genome.gov/sequencingcostsdata [Accessed 03/05/2018].
WHITWORTH, K. M., ROWLAND, R. R. R., EWEN, C. L., TRIBLE, B. R., KERRIGAN, M. A., CINO-OZUNA, A. G., SAMUEL, M. S., LIGHTNER, J. E., MCLAREN, D. G., MILEHAM, A. J., WELLS, K. D. & PRATHER, R. S. 2015. Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nature Biotechnology,* 34**,** 20.
WIGGANS, G. R., COOPER, T. A., VANRADEN, P. M., OLSON, K. M. & TOOKER, M. E. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J Dairy Sci,* 95**,** 1552-8.
WILLIAMS, A., STORTON, D., BUCKLES, J., LLINAS, M. & WANG, W. 2012. Improvement of PCR-free NGS Library Preparation to Obtain Uniform Read Coverage of Genome with Extremely High AT Content. *Journal of Biomolecular Techniques : JBT,* 23**,** S34-S34.
WINFIELD, M. O., WILKINSON, P. A., ALLEN, A. M., BARKER, G. L., COGHILL, J. A., BURRIDGE, A., HALL, A., BRENCHLEY, R. C., D'AMORE, R., HALL, N., BEVAN, M. W., RICHMOND, T., GERHARDT, D. J., JEDDELOH, J. A. & EDWARDS, K. J. 2012. Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol J,* 10**,** 733-42.
WITTKE-THOMPSON, J. K., PLUZHNIKOV, A. & COX, N. J. 2005. Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *American Journal of Human Genetics,* 76**,** 967-986.
WONG, K., KEANE, T. M., STALKER, J. & ADAMS, D. J. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biology,* 11**,** R128.
WORTHEY, E. A., MAYER, A. N., SYVERSON, G. D., HELBLING, D., BONACCI, B. B., DECKER, B., SERPE, J. M., DASU, T., TSCHANNEN, M. R., VEITH, R. L., BASEHORE, M. J., BROECKEL, U., TOMITA-MITCHELL, A., ARCA, M. J., CASPER, J. T., MARGOLIS, D. A., BICK, D. P., HESSNER, M. J., ROUTES, J. M., VERBSKY, J. W., JACOB, H. J. & DIMMOCK, D. P. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med,* 13**,** 255-62.
WRAGG, J. W., FINNITY, J. P., ANDERSON, J. A., FERGUSON, H. J., PORFIRI, E., BHATT, R. I., MURRAY, P. G., HEATH, V. L. & BICKNELL, R. 2016. MCAM and LAMA4 Are Highly Enriched in Tumor Blood Vessels of Renal Cell Carcinoma and Predict Patient Outcome. *Cancer Res,* 76**,** 2314-26.
WRIGHT, C. F., FITZGERALD, T. W., JONES, W. D., CLAYTON, S., MCRAE, J. F., VAN KOGELENBERG, M., KING, D. A., AMBRIDGE, K., BARRETT, D. M., BAYZETINOVA, T., BEVAN, A. P., BRAGIN, E., CHATZIMICHALI, E. A., GRIBBLE, S., JONES, P., KRISHNAPPA, N., MASON, L. E., MILLER, R., MORLEY, K. I., PARTHIBAN, V.,

PRIGMORE, E., RAJAN, D., SIFRIM, A., SWAMINATHAN, G. J., TIVEY, A. R., MIDDLETON, A., PARKER, M., CARTER, N. P., BARRETT, J. C., HURLES, M. E., FITZPATRICK, D. R. & FIRTH, H. V. 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet,* 385**,** 1305-1314.

WU, C. & SUN, D. 2015. GABA receptors in brain development, function, and injury. *Metab Brain Dis,* 30**,** 367-79.

WU, M., HENTZEL, M. D. & DZIUK, P. J. 1988. Effect of stage of gestation, litter size and uterine space on the incidence of mummified fetuses in pigs. *Journal of animal science,* 66 12**,** 3202-7.

WU, MICHAEL C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. & LIN, X. 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics,* 89**,** 82-93.

WU, S. H., SCHWARTZ, R. S., WINTER, D. J., CONRAD, D. F. & CARTWRIGHT, R. A. 2017. Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions. *Bioinformatics,* 33**,** 2322-2329.

WU, X. L., XU, J., FENG, G., WIGGANS, G. R., TAYLOR, J. F., HE, J., QIAN, C., QIU, J., SIMPSON, B., WALKER, J. & BAUCK, S. 2016. Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications. *PLoS One,* 11**,** e0161719.

XANTHOUDAKIS, S., SMEYNE, R. J., WALLACE, J. D. & CURRAN, T. 1996. The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice. *Proceedings of the National Academy of Sciences,* 93**,** 8919.

XI, R., HADJIPANAYIS, A. G., LUQUETTE, L. J., KIM, T.-M., LEE, E., ZHANG, J., JOHNSON, M. D., MUZNY, D. M., WHEELER, D. A., GIBBS, R. A., KUCHERLAPATI, R. & PARK, P. J. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences,* 108**,** E1128-E1136.

XU, S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics,* 163**,** 789-801.

XU, W. M., SHI, Q. X., CHEN, W. Y., ZHOU, C. X., NI, Y., ROWLANDS, D. K., YI LIU, G., ZHU, H., MA, Z. G., WANG, X. F., CHEN, Z. H., ZHOU, S. C., DONG, H. S., ZHANG, X. H., CHUNG, Y. W., YUAN, Y. Y., YANG, W. X. & CHAN, H. C. 2007. Cystic fibrosis transmembrane conductance regulator is vital to sperm fertilizing capacity and male fertility. *Proceedings of the National Academy of Sciences,* 104**,** 9816.

XU, Y., XIAO, B., JIANG, W. T., WANG, L., GEN, H. Q., CHEN, Y. W., SUN, Y. & JI, X. 2014. A novel mutation identified in PKHD1 by targeted exome sequencing: guiding prenatal diagnosis for an ARPKD family. *Gene,* 551**,** 33-8.

XUAN, J., YU, Y., QING, T., GUO, L. & SHI, L. 2013. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters,* 340**,** 284-295.

YALCIN, B., ADAMS, D. J., FLINT, J. & KEANE, T. M. 2012a. Next-generation sequencing of experimental mouse strains. *Mammalian genome : official journal of the International Mammalian Genome Society,* 23**,** 490-498.
YALCIN, B., WONG, K., BHOMRA, A., GOODSON, M., KEANE, T. M., ADAMS, D. J. & FLINT, J. 2012b. The fine-scale architecture of structural variants in 17 mouse genomes. *Genome biology,* 13**,** R18.
YAN, X. J., XU, J., GU, Z. H., PAN, C. M., LU, G., SHEN, Y., SHI, J. Y., ZHU, Y. M., TANG, L., ZHANG, X. W., LIANG, W. X., MI, J. Q., SONG, H. D., LI, K. Q., CHEN, Z. & CHEN, S. J. 2011. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat Genet,* 43**,** 309-15.
YANDELL, M., HUFF, C., HU, H., SINGLETON, M., MOORE, B., XING, J., JORDE, L. B. & REESE, M. G. 2011. A probabilistic disease-gene finder for personal genomes. *Genome Res,* 21**,** 1529-42.
YANG, B., ZHANG, W., ZHANG, Z., FAN, Y., XIE, X., AI, H., MA, J., XIAO, S., HUANG, L. & REN, J. 2013. Genome-Wide Association Analyses for Fatty Acid Composition in Porcine Muscle and Abdominal Fat Tissues. *PLOS ONE,* 8**,** e65554.
YANG, G.-D., YANG, X.-M., LU, H., REN, Y., MA, M.-Z., ZHU, L.-Y., WANG, J.-H., SONG, W.-W., ZHANG, W.-M., ZHANG, R. & ZHANG, Z.-G. 2014. SERPINA3 promotes endometrial cancer cells growth by regulating G2/M cell cycle checkpoint and apoptosis. *International Journal of Clinical and Experimental Pathology,* 7**,** 1348-1358.
YANG, L., GUELL, M., NIU, D., GEORGE, H., LESHA, E., GRISHIN, D., AACH, J., SHROCK, E., XU, W., POCI, J., CORTAZIO, R., WILKINSON, R. A., FISHMAN, J. A. & CHURCH, G. 2015. Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science,* 350**,** 1101-4.
YANG, R., FANG, S., WANG, J., ZHANG, C., ZHANG, R., LIU, D., ZHAO, Y., HU, X. & LI, N. 2017. Genome-wide analysis of structural variants reveals genetic differences in Chinese pigs. *PLOS ONE,* 12**,** e0186721.
YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics,* 13**,** 134-134.
YEUNG, T. L., LEUNG, C. S., WONG, K. K., SAMIMI, G., THOMPSON, M. S., LIU, J., ZAID, T. M., GHOSH, S., BIRRER, M. J. & MOK, S. C. 2013. TGF-beta modulates ovarian cancer invasion by upregulating CAF-derived versican in the tumor microenvironment. *Cancer Res,* 73**,** 5016-28.
YI, M., ZHAO, Y., JIA, L., HE, M., KEBEBEW, E. & STEPHENS, R. M. 2014. Performance comparison of SNP detection tools with illumina exome sequencing data--an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res,* 42**,** e101.

YNTEMA, H. G., OUDAKKER, A. R., KLEEFSTRA, T., HAMEL, B. C. J., VAN BOKHOVEN, H., CHELLY, J., KALSCHEUER, V. M., FRYNS, J.-P., RAYNAUD, M., MOIZARD, M.-P. & MORAINE, C. 2002. In-frame deletion in MECP2 causes mild nonspecific mental retardation. *American Journal of Medical Genetics,* 107**,** 81-83.

YU, X. & SUN, S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics,* 14**,** 1-15.

ZAHRADKA, K., SLADE, D., BAILONE, A., SOMMER, S., AVERBECK, D., PETRANOVIC, M., LINDNER, A. B. & RADMAN, M. 2006. Reassembly of shattered chromosomes in Deinococcus radiodurans. *Nature,* 443**,** 569.

ZAK, L. J., GAUSTAD, A. H., BOLARIN, A., BROEKHUIJSE, M., WALLING, G. A. & KNOL, E. F. 2017. Genetic control of complex traits, with a focus on reproduction in pigs. *Mol Reprod Dev,* 84**,** 1004-1011.

ZERBINO, D. R., ACHUTHAN, P., AKANNI, W., AMODE, M R., BARRELL, D., BHAI, J., BILLIS, K., CUMMINS, C., GALL, A., GIRÓN, C. G., GIL, L., GORDON, L., HAGGERTY, L., HASKELL, E., HOURLIER, T., IZUOGU, O. G., JANACEK, S. H., JUETTEMANN, T., TO, J. K., LAIRD, M. R., LAVIDAS, I., LIU, Z., LOVELAND, J. E., MAUREL, T., MCLAREN, W., MOORE, B., MUDGE, J., MURPHY, D. N., NEWMAN, V., NUHN, M., OGEH, D., ONG, C. K., PARKER, A., PATRICIO, M., RIAT, H. S., SCHUILENBURG, H., SHEPPARD, D., SPARROW, H., TAYLOR, K., THORMANN, A., VULLO, A., WALTS, B., ZADISSA, A., FRANKISH, A., HUNT, S. E., KOSTADIMA, M., LANGRIDGE, N., MARTIN, F. J., MUFFATO, M., PERRY, E., RUFFIER, M., STAINES, D. M., TREVANION, S. J., AKEN, B. L., CUNNINGHAM, F., YATES, A. & FLICEK, P. 2018. Ensembl 2018. *Nucleic Acids Research,* 46**,** D754-D761.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res,* 18.

ZHANG, G. 2015. Bird sequencing project takes off. *Nature,* 522**,** 34.

ZHANG, Q. & BACKSTROM, N. 2014. Assembly errors cause false tandem duplicate regions in the chicken (Gallus gallus) genome sequence. *Chromosoma,* 123**,** 165-8.

ZHANG, W., NG, H. W., SHU, M., LUO, H., SU, Z., GE, W., PERKINS, R., TONG, W. & HONG, H. 2015. Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. *J Genet,* 94**,** 731-40.

ZHENG, G. X. Y., LAU, B. T., SCHNALL-LEVIN, M., JAROSZ, M., BELL, J. M., HINDSON, C. M., KYRIAZOPOULOU-PANAGIOTOPOULOU, S., MASQUELIER, D. A., MERRILL, L., TERRY, J. M., MUDIVARTI, P. A., WYATT, P. W., BHARADWAJ, R., MAKAREWICZ, A. J., LI, Y., BELGRADER, P., PRICE, A. D., LOWE, A. J., MARKS, P., VURENS, G. M., HARDENBOL, P., MONTESCLAROS, L., LUO, M., GREENFIELD, L., WONG, A., BIRCH, D. E., SHORT, S. W., BJORNSON, K. P., PATEL, P., HOPMANS, E. S., WOOD, C., KAUR, S., LOCKWOOD, G. K., STAFFORD, D., DELANEY, J. P., WU, I., ORDONEZ, H. S., GRIMES, S. M., GREER, S., LEE, J. Y.,

BELHOCINE, K., GIORDA, K. M., HEATON, W. H., MCDERMOTT, G. P., BENT, Z. W., MESCHI, F., KONDOV, N. O., WILSON, R., BERNATE, J. A., GAUBY, S., KINDWALL, A., BERMEJO, C., FEHR, A. N., CHAN, A., SAXONOV, S., NESS, K. D., HINDSON, B. J. & JI, H. P. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology,* 34**,** 303.

ZHENG, H.-F., RONG, J.-J., LIU, M., HAN, F., ZHANG, X.-W., RICHARDS, J. B. & WANG, L. 2015. Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. *PLoS ONE,* 10**,** e0116487.

ZHOU, L. & HOLLIDAY, J. A. 2012. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *BMC Genomics,* 13**,** 703-703.

ZHU, Y., PASZTY, C., TURETSKY, T., TSAI, S., KUYPERS, F. A., LEE, G., COOPER, P., GALLAGHER, P. G., STEVENS, M. E., RUBIN, E., MOHANDAS, N. & MENTZER, W. C. 1999. Stomatocytosis Is Absent in “Stomatin”-Deficient Murine Red Blood Cells. *Blood,* 93**,** 2404-2410.